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.

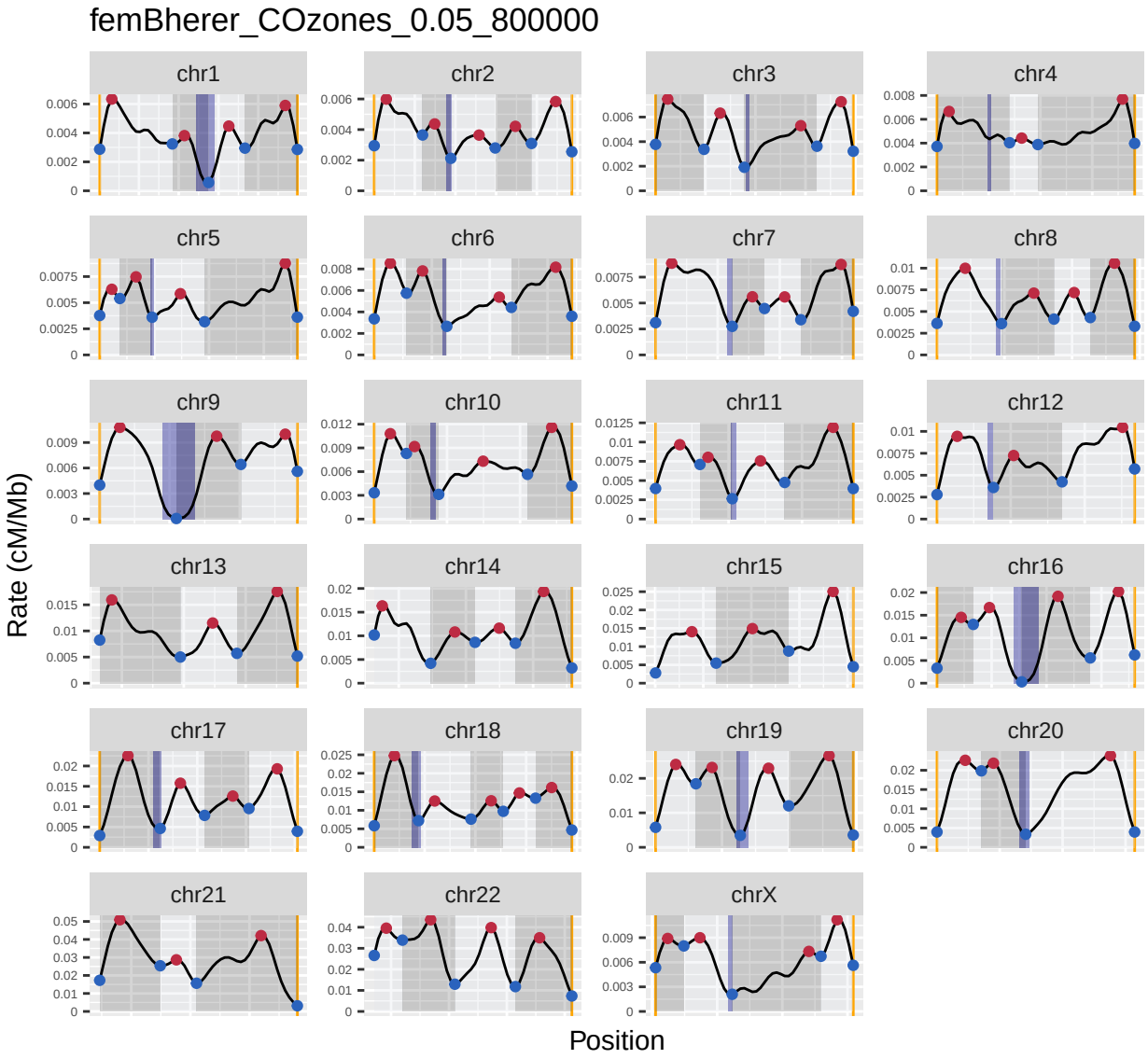


Figure 1: Crossover zones; centromeres in blue, workspace limits in orange.

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.

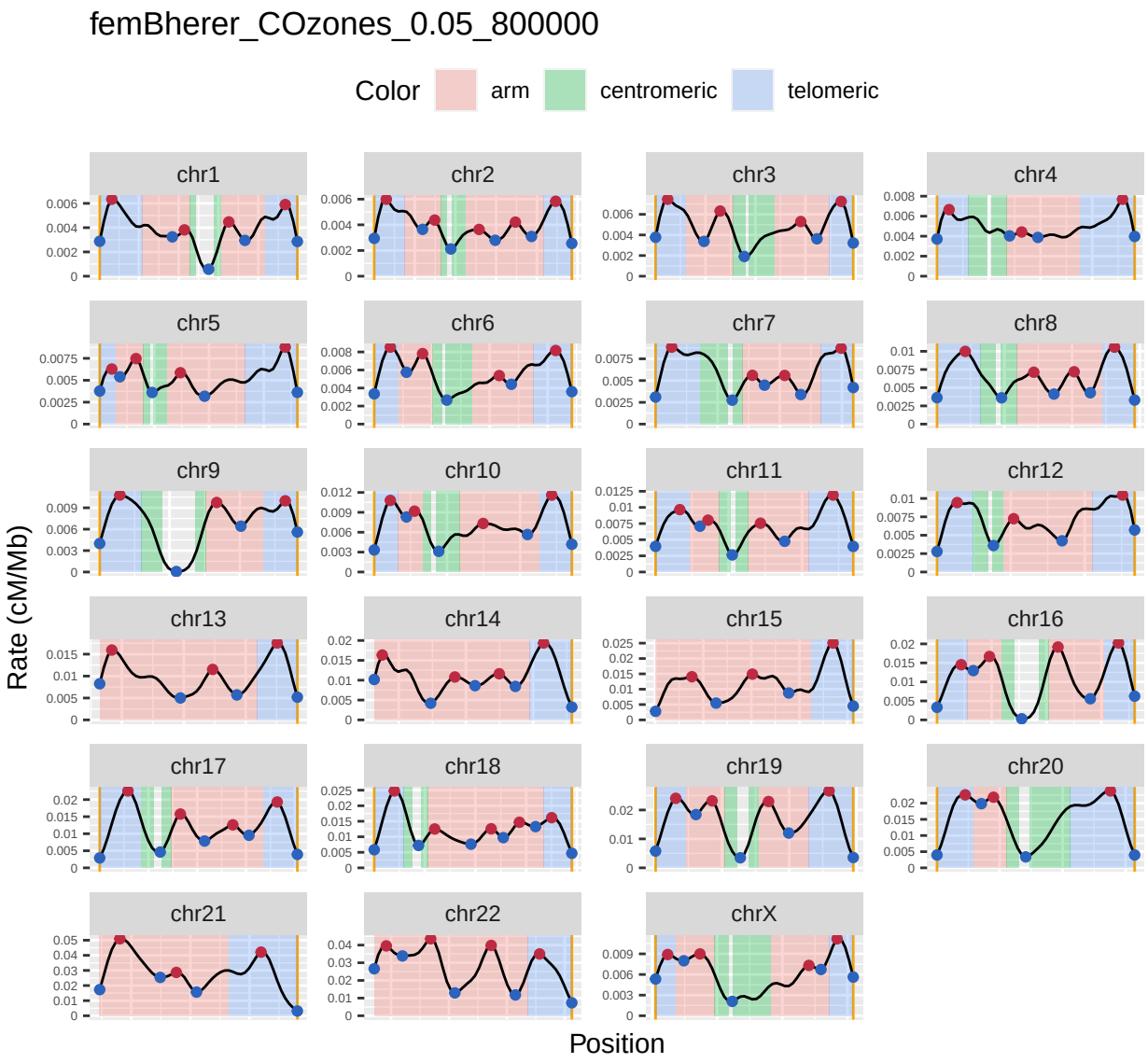
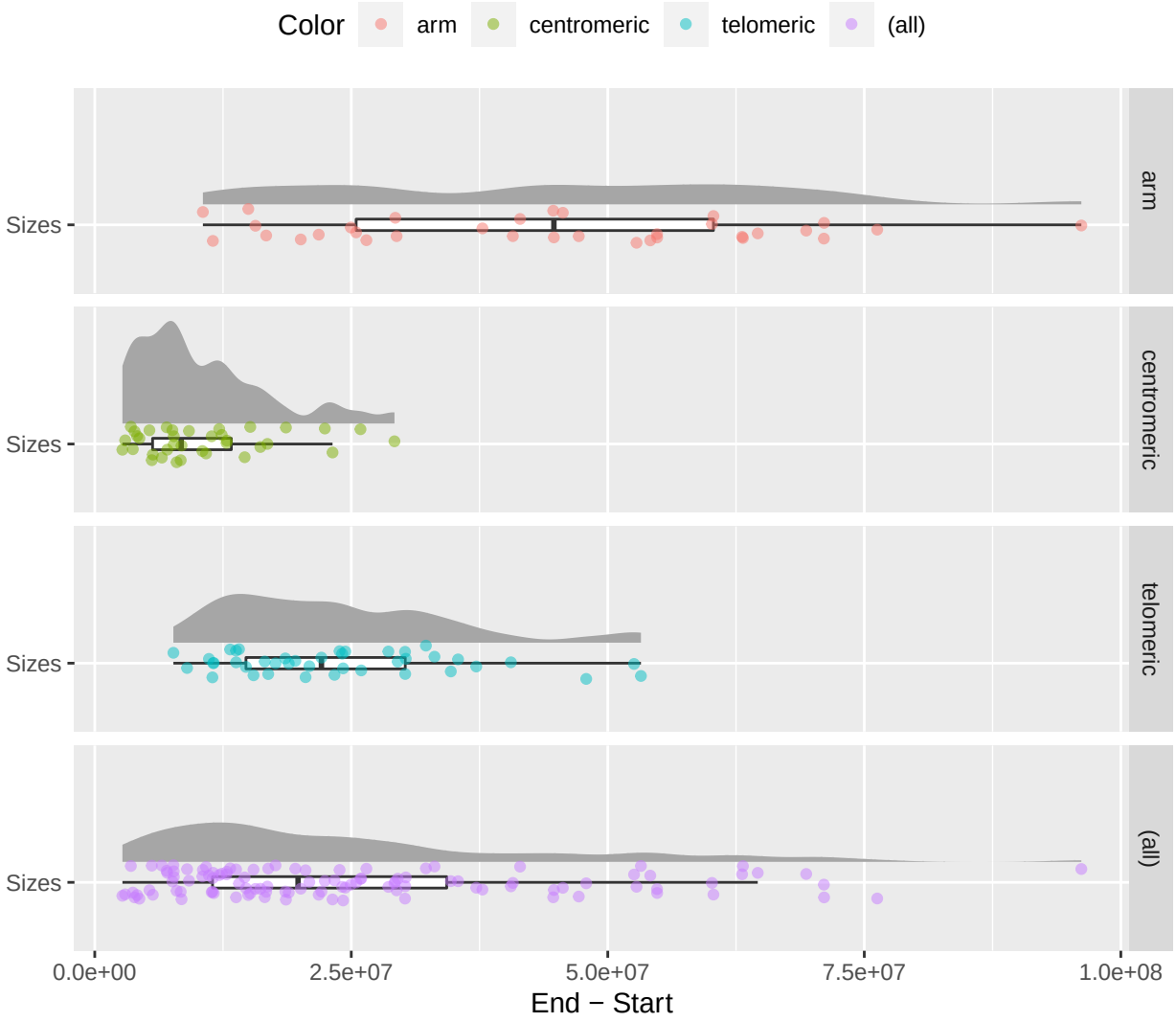
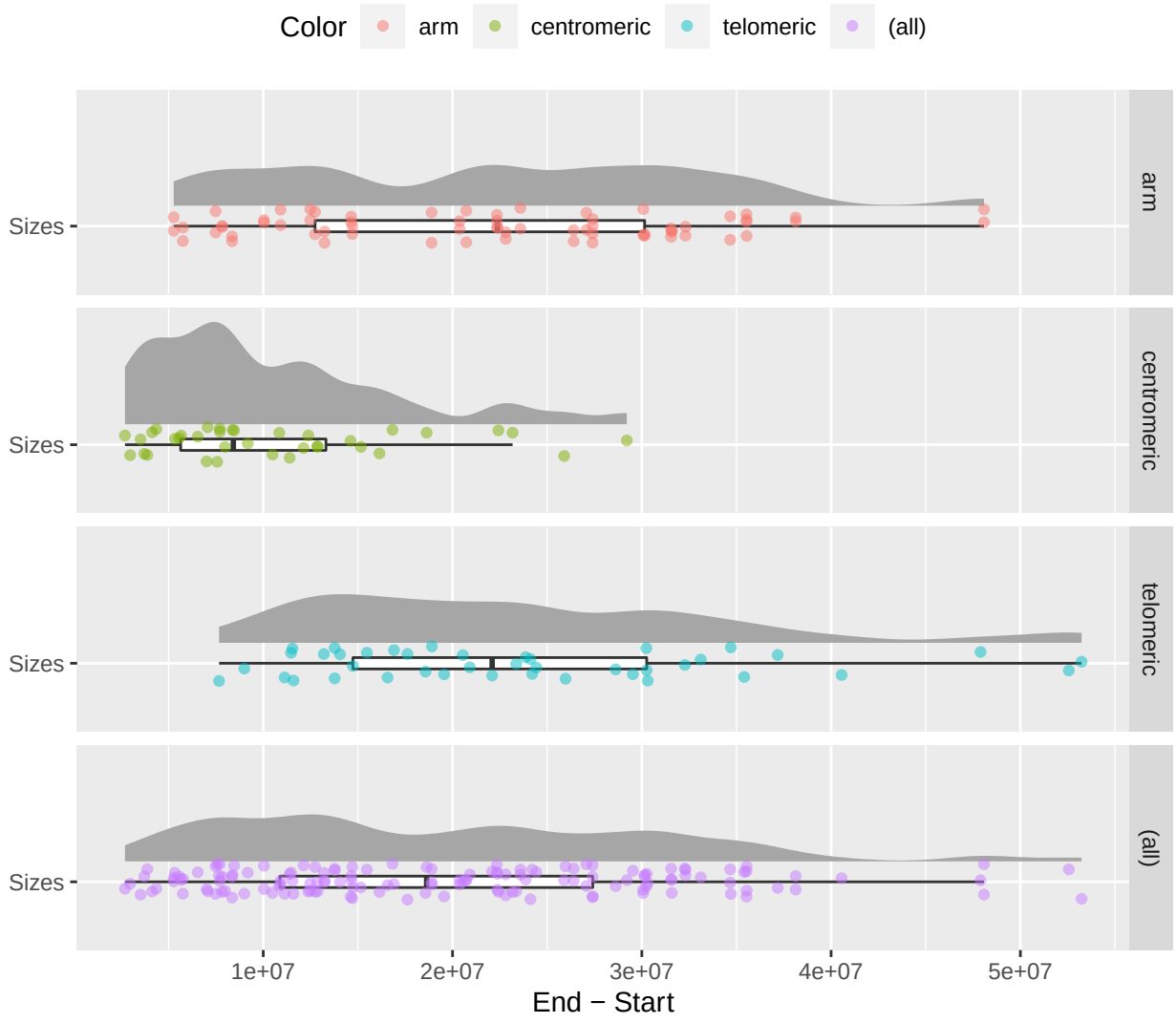


Figure 2: Color-coded windows for telomeric, centromeric and arm categories.

Distribution



Distribution



Numerical categories

Descriptive statistics

Raw data:

Chromosome	Start	End	Color	invCenter	NHCenter	NAHRCenter	Length.Mb	RepCounts	log10RepCounts	WAvgRate.perMb	ChrType
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	113381273	135473442	telomeric	1	1	0	22.092163	170	2.230449	2.1846155	A
chr10	42436305	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.

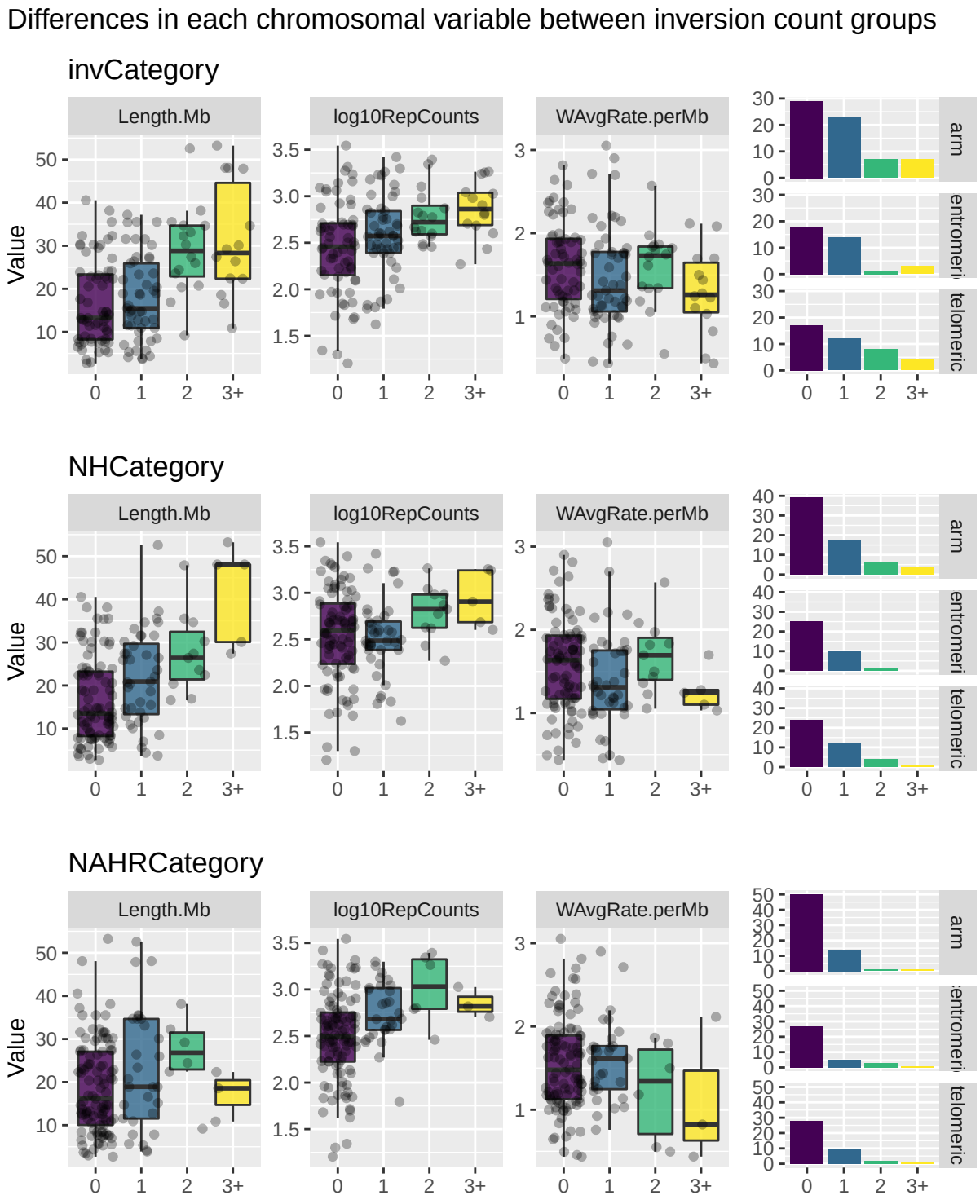


Figure 4: Potential effect of independent variables on the different types of invesions.

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

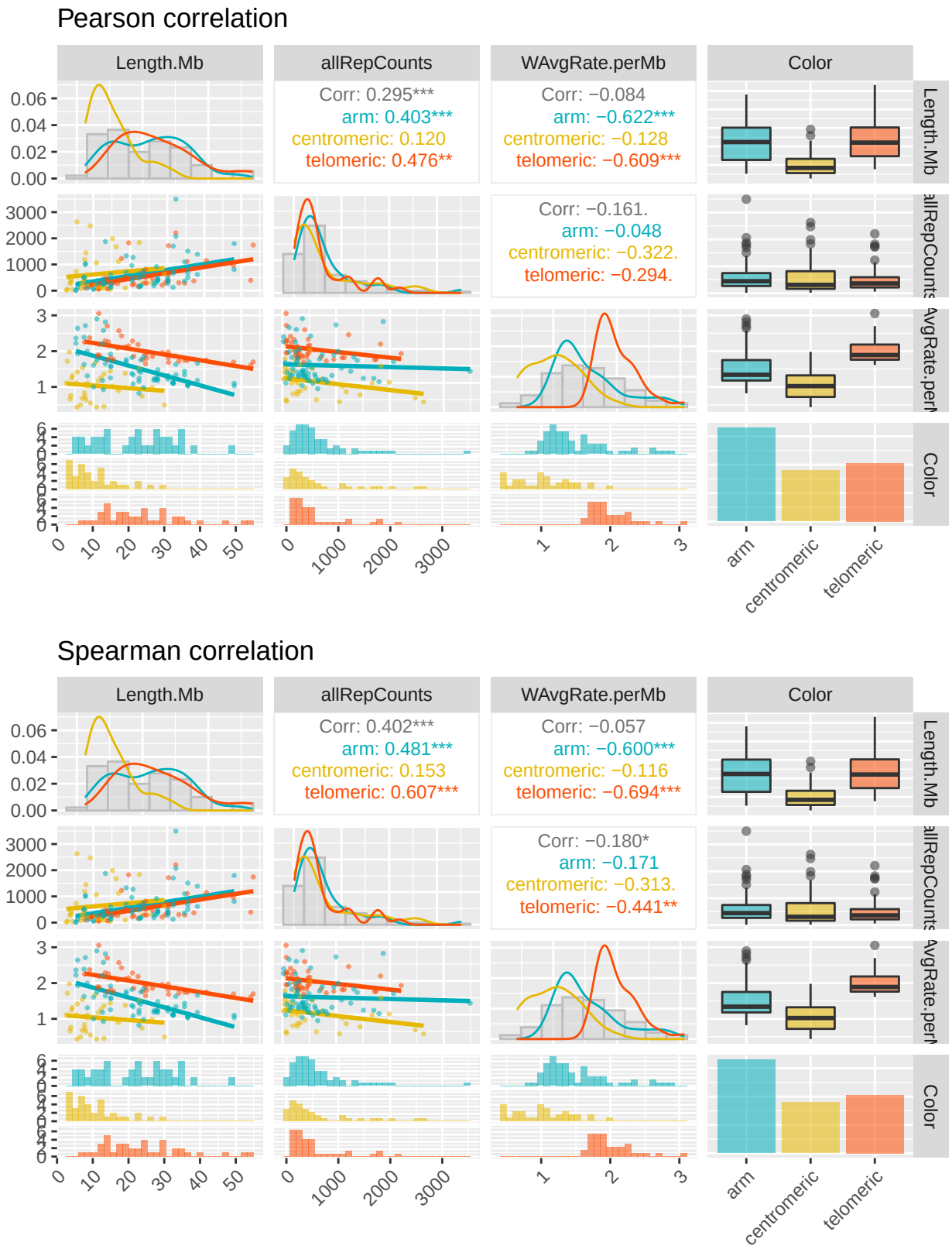


Figure 5: Correlations between variables.

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         0.0749455  0.0204365  3.6672
## allRepCounts      0.0003626  0.0003049  1.1894
## Colorcentromeric  0.4171659  0.5636414  0.7401
## Colortelomeric    0.1992110  0.4609093  0.4322
## WAvgRate.perMb   -0.1982014  0.4675271 -0.4239
## ChromTypeX        2.0128414  0.7766110  2.5918
##
## Intercepts:
##      Value   Std. Error t value
## 0|1   1.3059  1.0358     1.2608
## 1|2   3.1278  1.0644     2.9385
## 2|3+  4.2464  1.1143     3.8110
##
## Residual Deviance: 309.0743
## AIC: 327.0743
```

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

	Value	Std. Error	t value	p value
Length.Mb	0.0749455	0.0204365	3.6672482	0.0002452
allRepCounts	0.0003626	0.0003049	1.1894366	0.2342679
Colorcentromeric	0.4171659	0.5636414	0.7401265	0.4592233
Colortelomeric	0.1992110	0.4609093	0.4322129	0.6655867
WAvgRate.perMb	-0.1982014	0.4675271	-0.4239355	0.6716128
ChromTypeX	2.0128414	0.7766110	2.5918269	0.0095468
0|1	1.3059011	1.0358008	1.2607648	0.2073936
1|2	3.1277635	1.0644248	2.9384541	0.0032985
2|3+	4.2464257	1.1142694	3.8109506	0.0001384

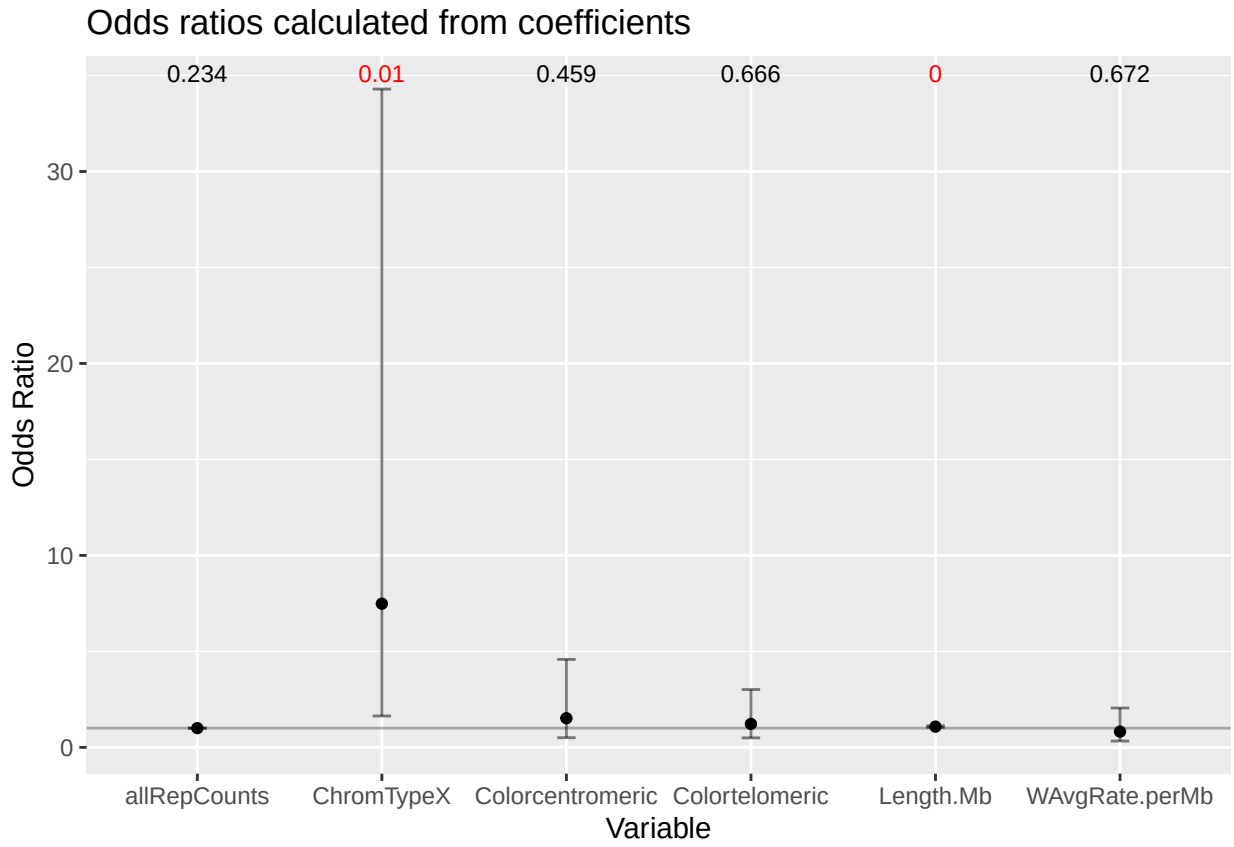
We can also get confidence intervals for the parameter estimates. These can be obtained either by profiling the likelihood function or by using the standard errors and assuming a normal distribution. Note that profiled CIs are not symmetric (although they are usually close to symmetric). If the 95% CI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

	2.5 %	97.5 %
Length.Mb	0.0348908	0.1150002
allRepCounts	-0.0002349	0.0009602
Colorcentromeric	-0.6875509	1.5218828
Colortelomeric	-0.7041547	1.1025767
WAvgRate.perMb	-1.1145377	0.7181350
ChromTypeX	0.4907117	3.5349710

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	1.0778254	1.0355067	1.121874
allRepCounts	1.0003627	0.9997651	1.000961
Colorcentromeric	1.5176543	0.5028060	4.580842
Colortelomeric	1.2204394	0.4945264	3.011917
WAvgRate.perMb	0.8202047	0.3280669	2.050605
ChromTypeX	7.4845534	1.6334784	34.294019

Example of interpretation: "For 1 unit increase in Length.Mb, a window is 1.0778254 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.

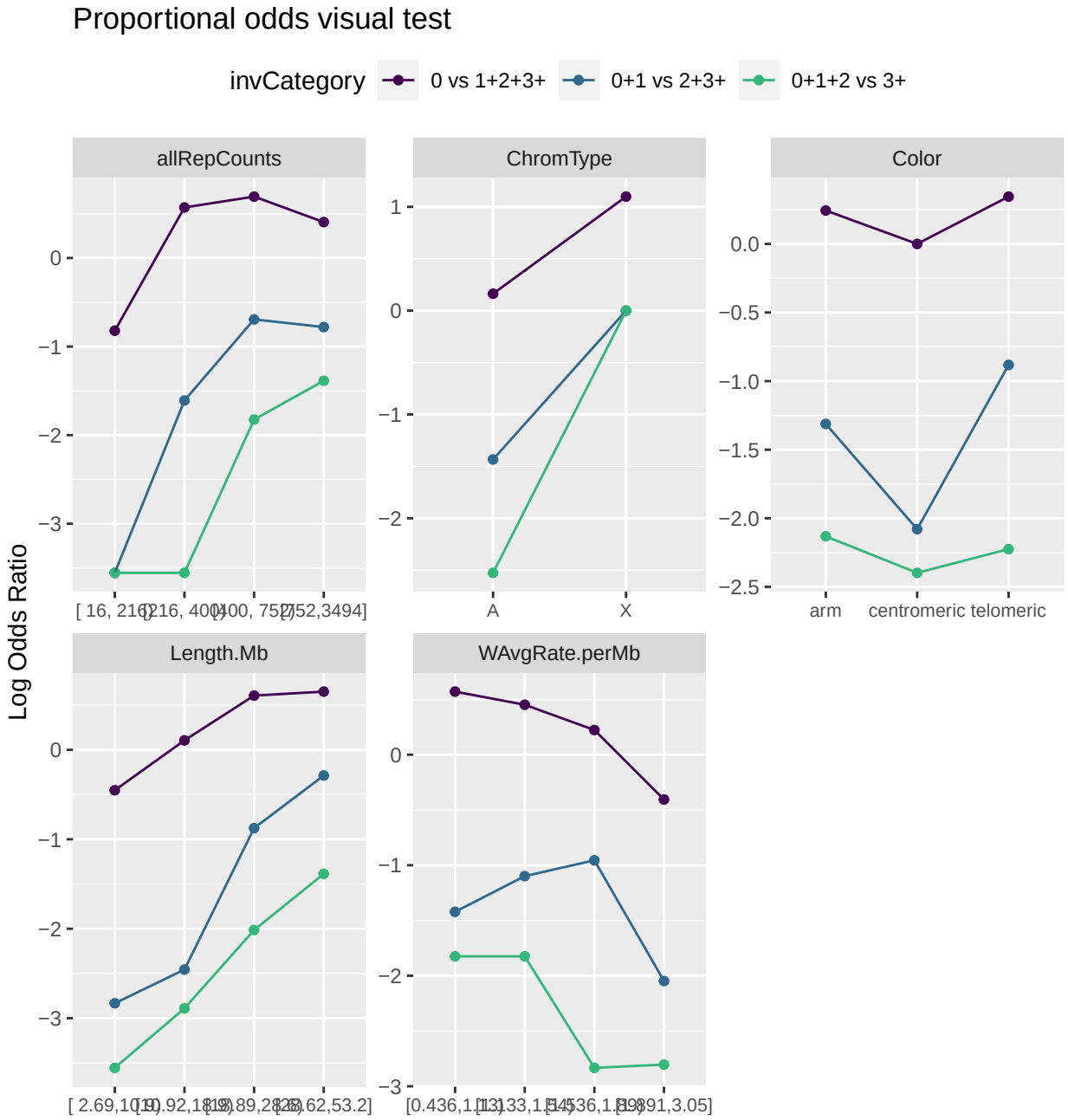
We test the parallel regression assumption with a Brant test:

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

```
## ------------------------------------------------------
## Test for     X2  df  probability
## ------------------------------------------------------
## Omnibus          62.99   12  0
## Length.Mb        9.72    2   0.01
## allRepCounts     0   2   1
## Colorcentromeric 2.1 2   0.35
## Colortelomeric       1.06    2   0.59
## WAvgRate.perMb       0.67    2   0.71
## ChromTypeX       9.99    2   0.01
## ------------------------------------------------------
##
## H0: Parallel Regression Assumption holds
```

	X2	df	probability
Omnibus	62.9898415	12	0.0000000
Length.Mb	9.7168082	2	0.0077629
allRepCounts	0.0029805	2	0.9985108
Colorcentromeric	2.0999201	2	0.3499517
Colortelomeric	1.0607528	2	0.5883834
WAvgRate.perMb	0.6717577	2	0.7147097
ChromTypeX	9.9910725	2	0.0067681

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.

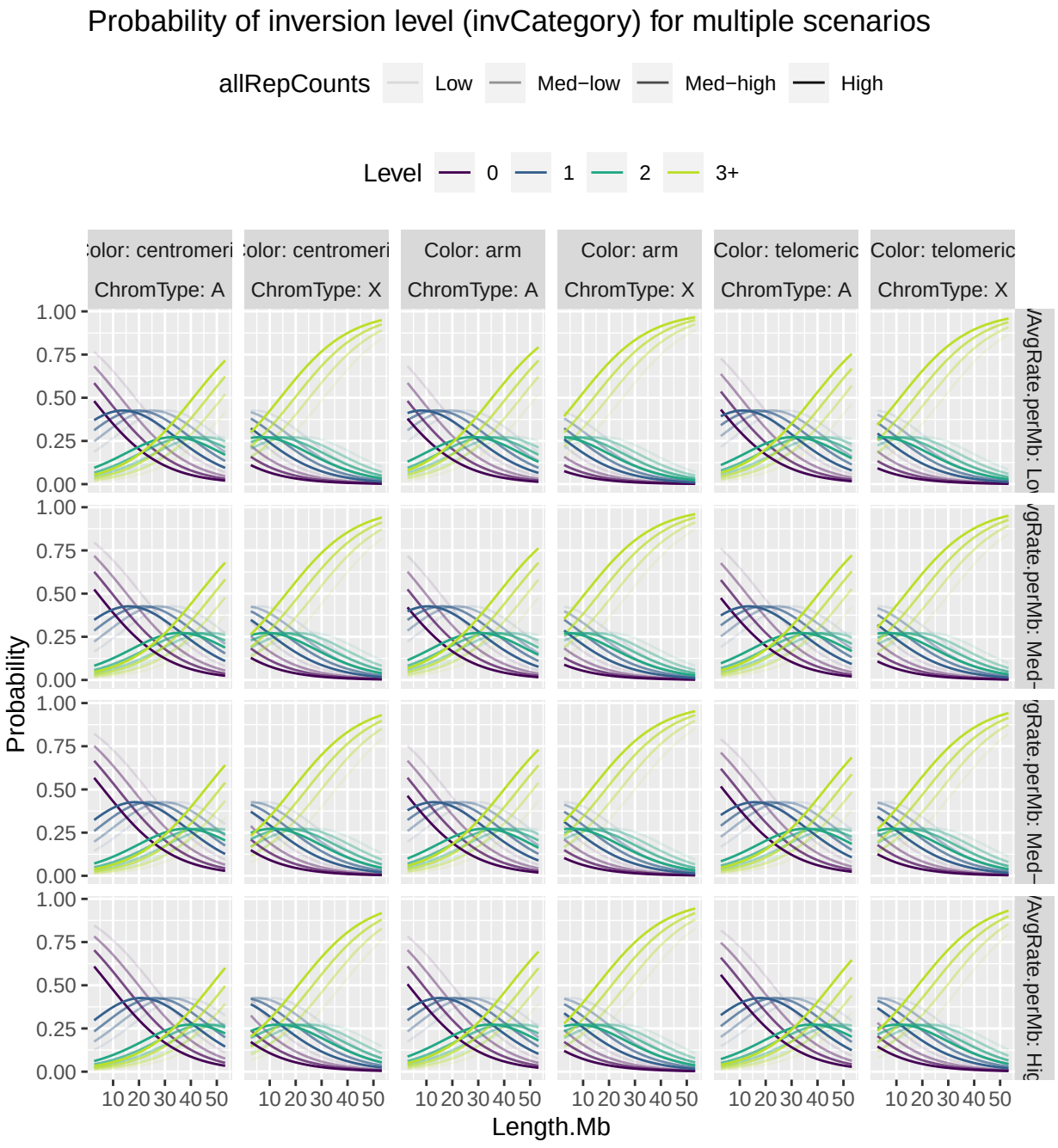


Figure 6: Probabiilty of having 0 to >3 inversions depending on multiple independent variables

NH inversions model

Model fitting

```
## Call:
## polr(formula = myFormula, data = winRegions, Hess = T)
##
## Coefficients:
##                       Value Std. Error t value
## Length.Mb         0.0883776  0.0226326  3.9049
## allRepCounts     -0.0003229  0.0003477 -0.9287
## Colorcentromeric  0.2288804  0.6141147  0.3727
## Colortelomeric   -0.1214429  0.5273087 -0.2303
## WAvgRate.perMb   -0.3173525  0.5595591 -0.5671
## ChromTypeX       -0.8588464  0.8687749 -0.9886
##
## Intercepts:
##      Value   Std. Error t value
## 0|1   1.5512  1.1771     1.3178
## 1|2   3.4035  1.2183     2.7936
## 2|3+  4.7942  1.3038     3.6770
##
## Residual Deviance: 248.1198
## AIC: 266.1198
```

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

	Value	Std. Error	t value	p value
Length.Mb	0.0883776	0.0226326	3.9048792	0.0000943
allRepCounts	-0.0003229	0.0003477	-0.9287264	0.3530309
Colorcentromeric	0.2288804	0.6141147	0.3726998	0.7093719
Colortelomeric	-0.1214429	0.5273087	-0.2303070	0.8178532
WAvgRate.perMb	-0.3173525	0.5595591	-0.5671474	0.5706140
ChromTypeX	-0.8588464	0.8687749	-0.9885718	0.3228727
0|1	1.5511854	1.1771451	1.3177521	0.1875866
1|2	3.4034597	1.2183253	2.7935559	0.0052132
2|3+	4.7941591	1.3038237	3.6769994	0.0002360

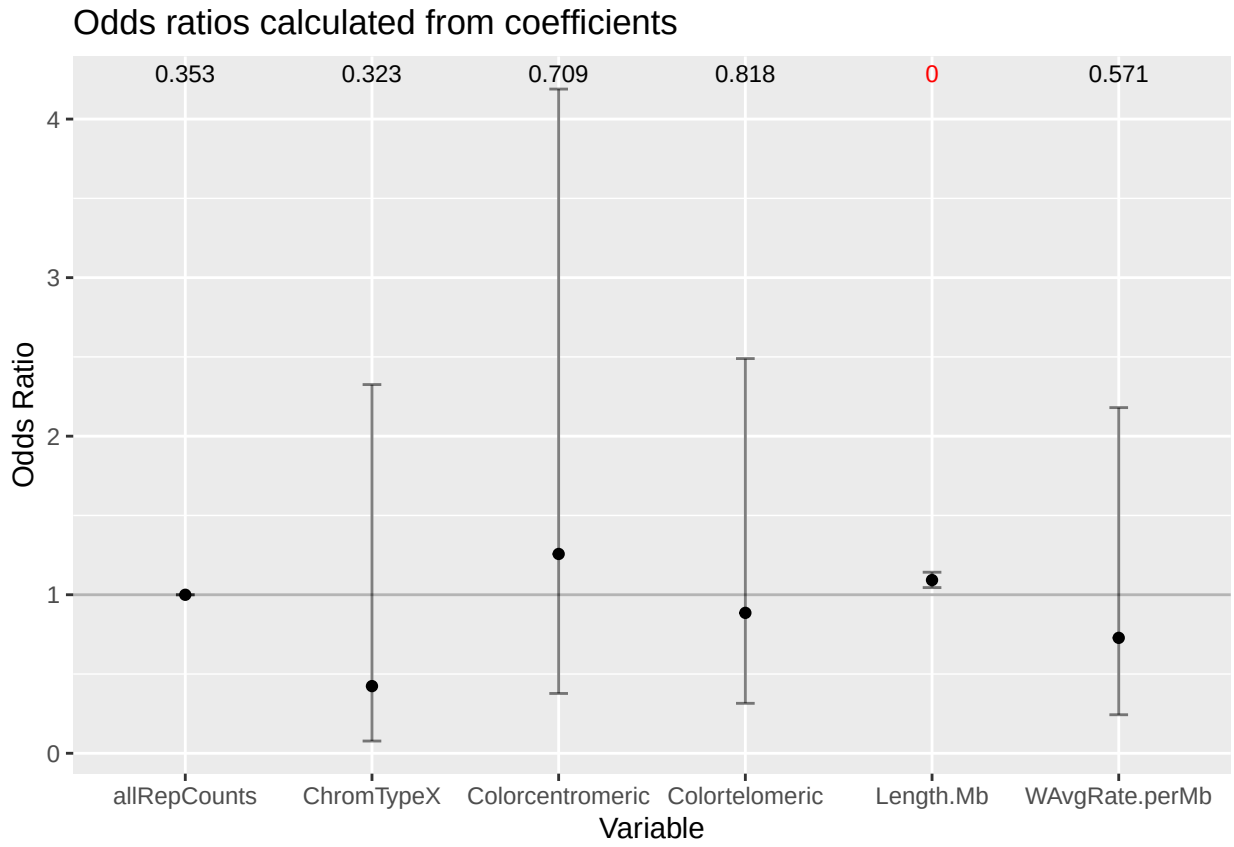
We can also get confidence intervals for the parameter estimates. These can be obtained either by profiling the likelihood function or by using the standard errors and assuming a normal distribution. Note that profiled CIs are not symmetric (although they are usually close to symmetric). If the 95% CI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

	2.5 %	97.5 %
Length.Mb	0.0440185	0.1327367
allRepCounts	-0.0010043	0.0003585
Colorcentromeric	-0.9747623	1.4325231
Colortelomeric	-1.1549489	0.9120632
WAvgRate.perMb	-1.4140682	0.7793632
ChromTypeX	-2.5616140	0.8439212

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	1.0924005	1.0450017	1.141949
allRepCounts	0.9996772	0.9989962	1.000359
Colorcentromeric	1.2571917	0.3772820	4.189256
Colortelomeric	0.8856417	0.3150736	2.489453
WAvgRate.perMb	0.7280740	0.2431521	2.180083
ChromTypeX	0.4236505	0.0771801	2.325468

Example of interpretation: "For 1 unit increase in Length.Mb, a window is 1.0924005 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.

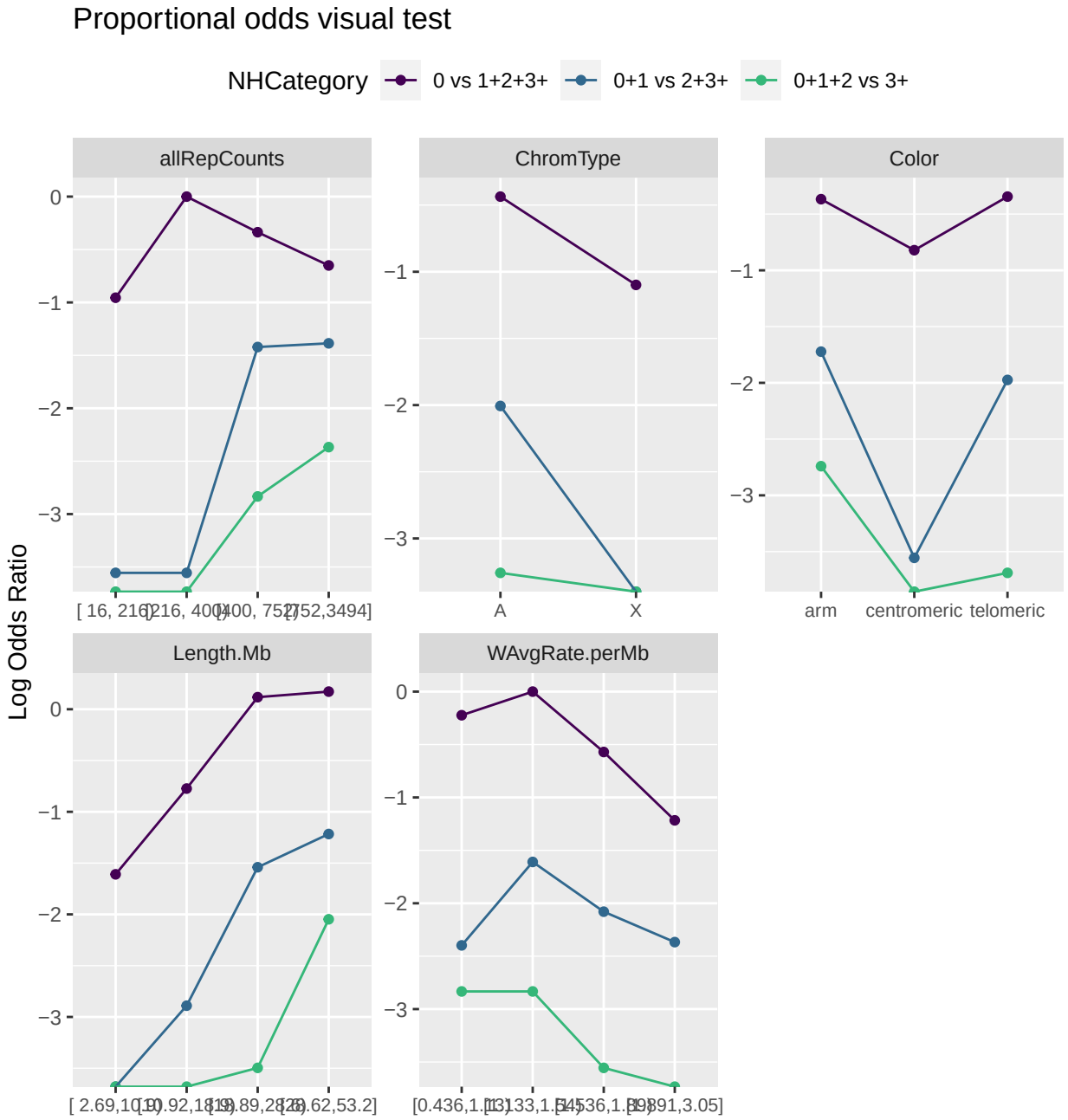
We test the parallel regression assumption with a Brant test:

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

```
## ------------------------------------------------------
## Test for     X2  df  probability
## ------------------------------------------------------
## Omnibus          19.16   12  0.08
## Length.Mb        3.58    2   0.17
## allRepCounts     1.09    2   0.58
## Colorcentromeric 0.12    2   0.94
## Colortelomeric       4.89    2   0.09
## WAvgRate.perMb       8.45    2   0.01
## ChromTypeX       0   2   1
## ------------------------------------------------------
##
## H0: Parallel Regression Assumption holds
```

	X2	df	probability
Omnibus	19.1612073	12	0.0847116
Length.Mb	3.5779305	2	0.1671330
allRepCounts	1.0895693	2	0.5799667
Colorcentromeric	0.1183776	2	0.9425288
Colortelomeric	4.8901861	2	0.0867181
WAvgRate.perMb	8.4547485	2	0.0145907
ChromTypeX	0.0001120	2	0.9999440

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.

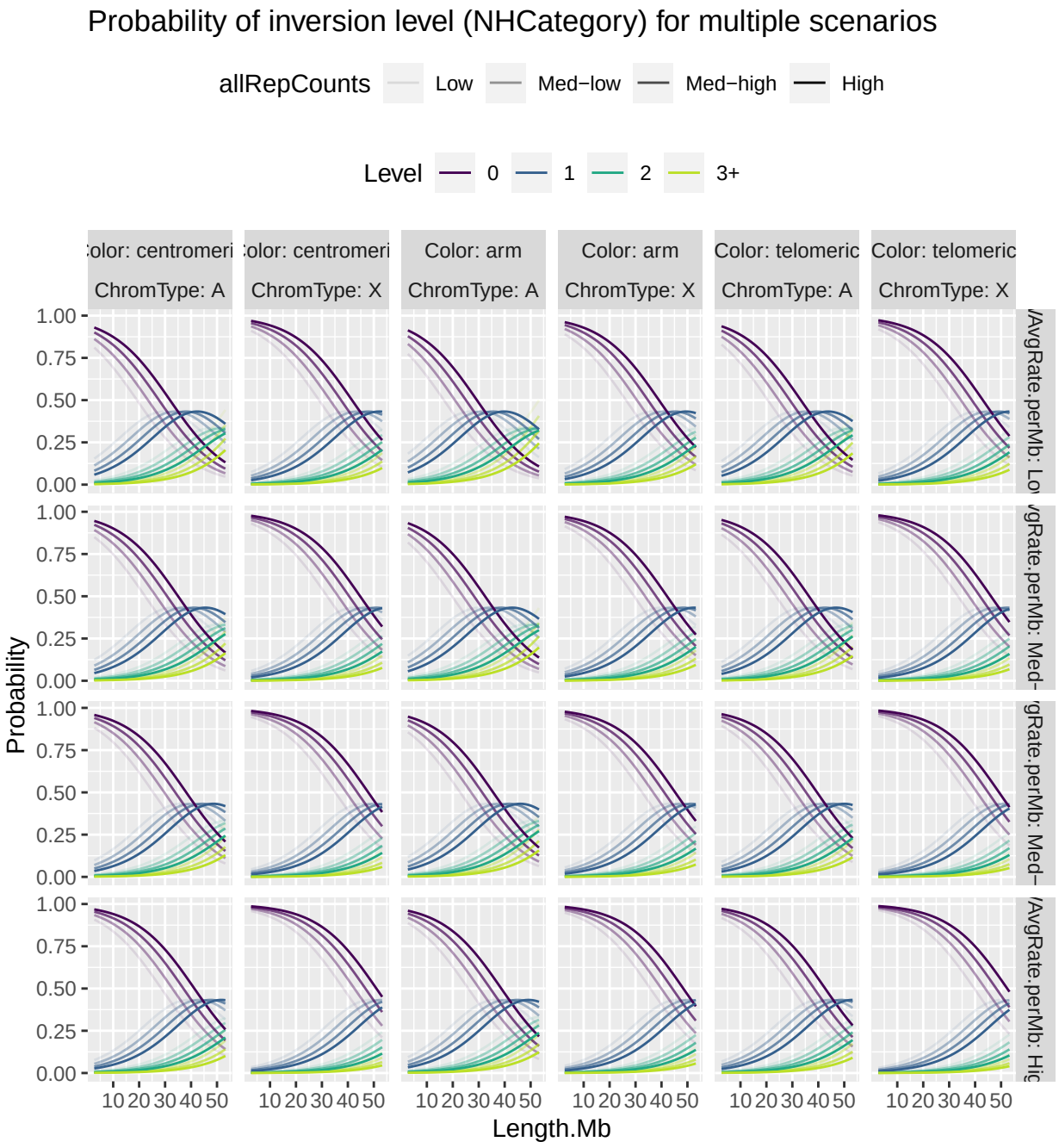


Figure 7: Probabiilty of having 0 to >3 inversions depending on multiple independent variables

NAHR inversions model

Model fitting

```
## Call:
## polr(formula = myFormula, data = winRegions, Hess = T)
##
## Coefficients:
##                      Value Std. Error t value
## Length.Mb        0.0244063  0.0237105  1.0293
## allRepCounts     0.0008745  0.0003637  2.4042
## Colorcentromeric 0.6056516  0.7237776  0.8368
## Colortelomeric   0.5515899  0.5585025  0.9876
## WAvgRate.perMb   0.0936625  0.6008008  0.1559
## ChromTypeX       3.1486795  0.8543621  3.6854
##
## Intercepts:
##      Value  Std. Error t value
## 0|1  2.7070 1.3429     2.0158
## 1|2  4.6541 1.4054     3.3116
## 2|3+ 6.0672 1.5103     4.0173
##
## Residual Deviance: 194.8853
## AIC: 212.8853
```

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

	Value	Std. Error	t value	p value
Length.Mb	0.0244063	0.0237105	1.0293454	0.3033174
allRepCounts	0.0008745	0.0003637	2.4041851	0.0162086
Colorcentromeric	0.6056516	0.7237776	0.8367924	0.4027093
Colortelomeric	0.5515899	0.5585025	0.9876229	0.3233374
WAvgRate.perMb	0.0936625	0.6008008	0.1558961	0.8761149
ChromTypeX	3.1486795	0.8543621	3.6854158	0.0002283
0|1	2.7069817	1.3429124	2.0157545	0.0438256
1|2	4.6541006	1.4054008	3.3115825	0.0009277
2|3+	6.0672459	1.5102743	4.0173138	0.0000589

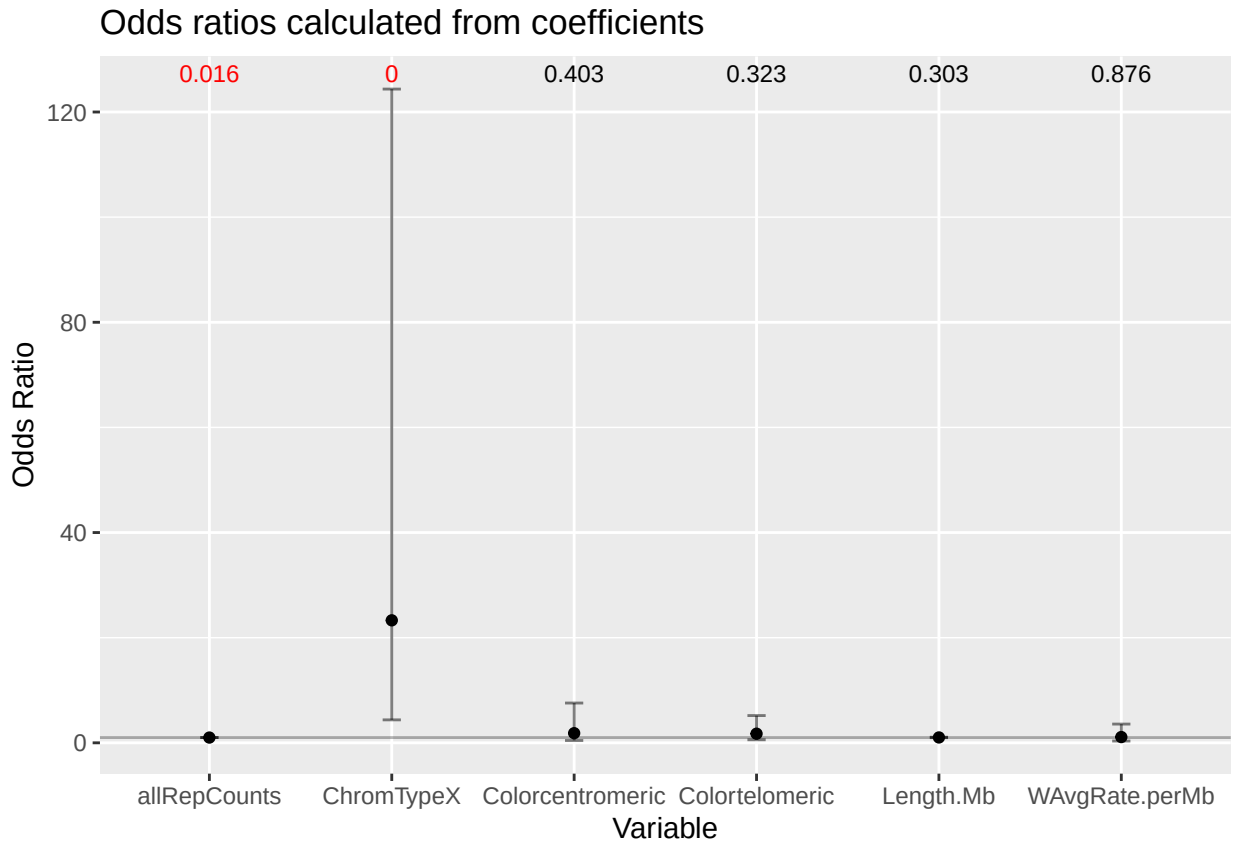
We can also get confidence intervals for the parameter estimates. These can be obtained either by profiling the likelihood function or by using the standard errors and assuming a normal distribution. Note that profiled CIs are not symmetric (although they are usually close to symmetric). If the 95% CI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

	2.5 %	97.5 %
Length.Mb	-0.0220654	0.0708779
allRepCounts	0.0001616	0.0015873
Colorcentromeric	-0.8129264	2.0242295
Colortelomeric	-0.5430549	1.6462348
WAvgRate.perMb	-1.0838854	1.2712104
ChromTypeX	1.4741606	4.8231984

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	1.024706	0.9781763	1.073450
allRepCounts	1.000875	1.0001616	1.001589
Colorcentromeric	1.832446	0.4435581	7.570276
Colortelomeric	1.736011	0.5809707	5.187411
WAvgRate.perMb	1.098189	0.3382786	3.565165
ChromTypeX	23.305270	4.3673682	124.362216

Example of interpretation: "For 1 unit increase in Length.Mb, a window is 1.0247065 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.

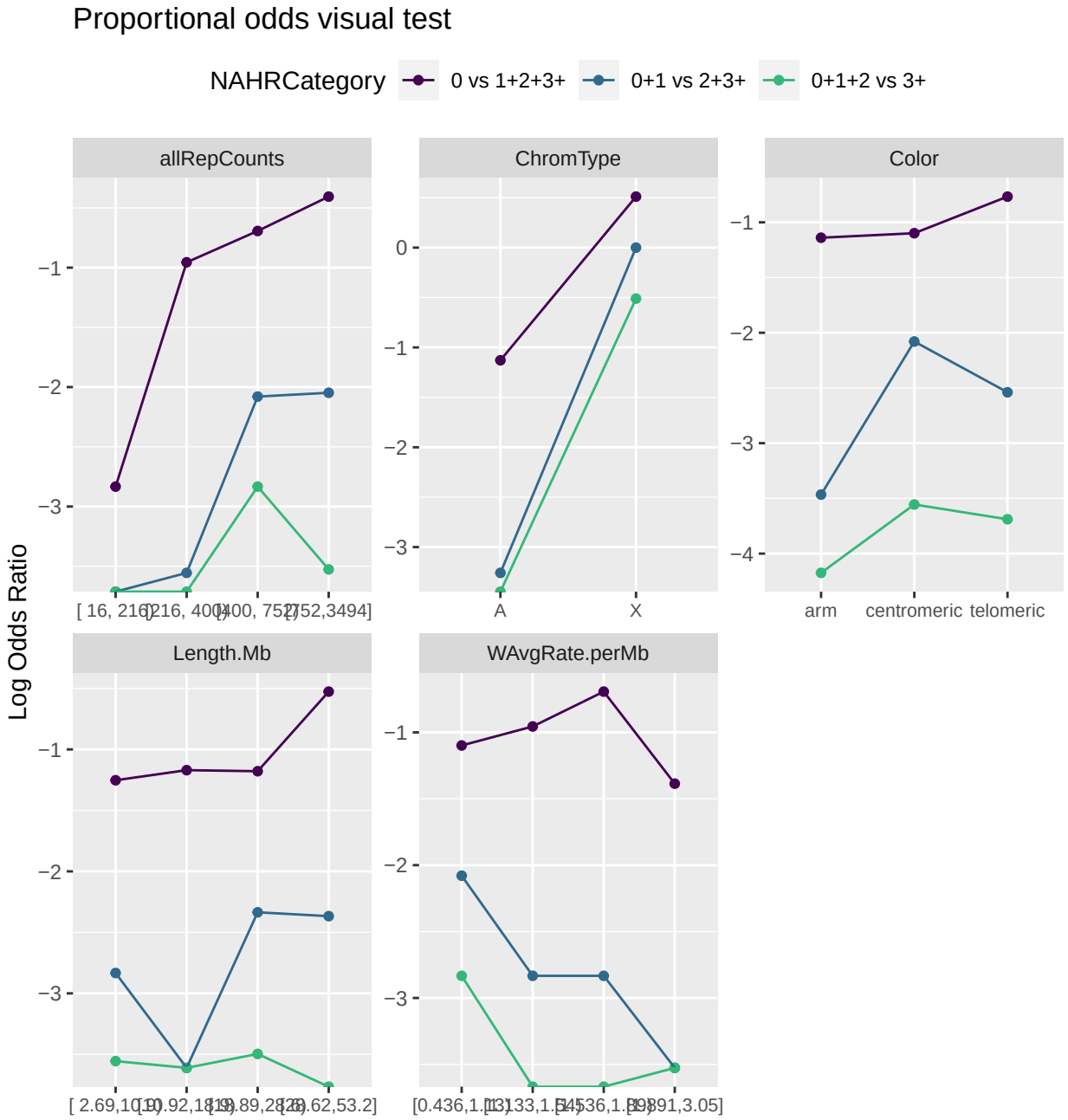
We test the parallel regression assumption with a Brant test:

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

```
## Error in solve.default(D %*% varBeta %*% t(D)): system is computationally singular: reciprocal condi
```

	X2	df	probability
Omnibus	19.1612073	12	0.0847116
Length.Mb	3.5779305	2	0.1671330
allRepCounts	1.0895693	2	0.5799667
Colorcentromeric	0.1183776	2	0.9425288
Colortelomeric	4.8901861	2	0.0867181
WAvgRate.perMb	8.4547485	2	0.0145907
ChromTypeX	0.0001120	2	0.9999440

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.

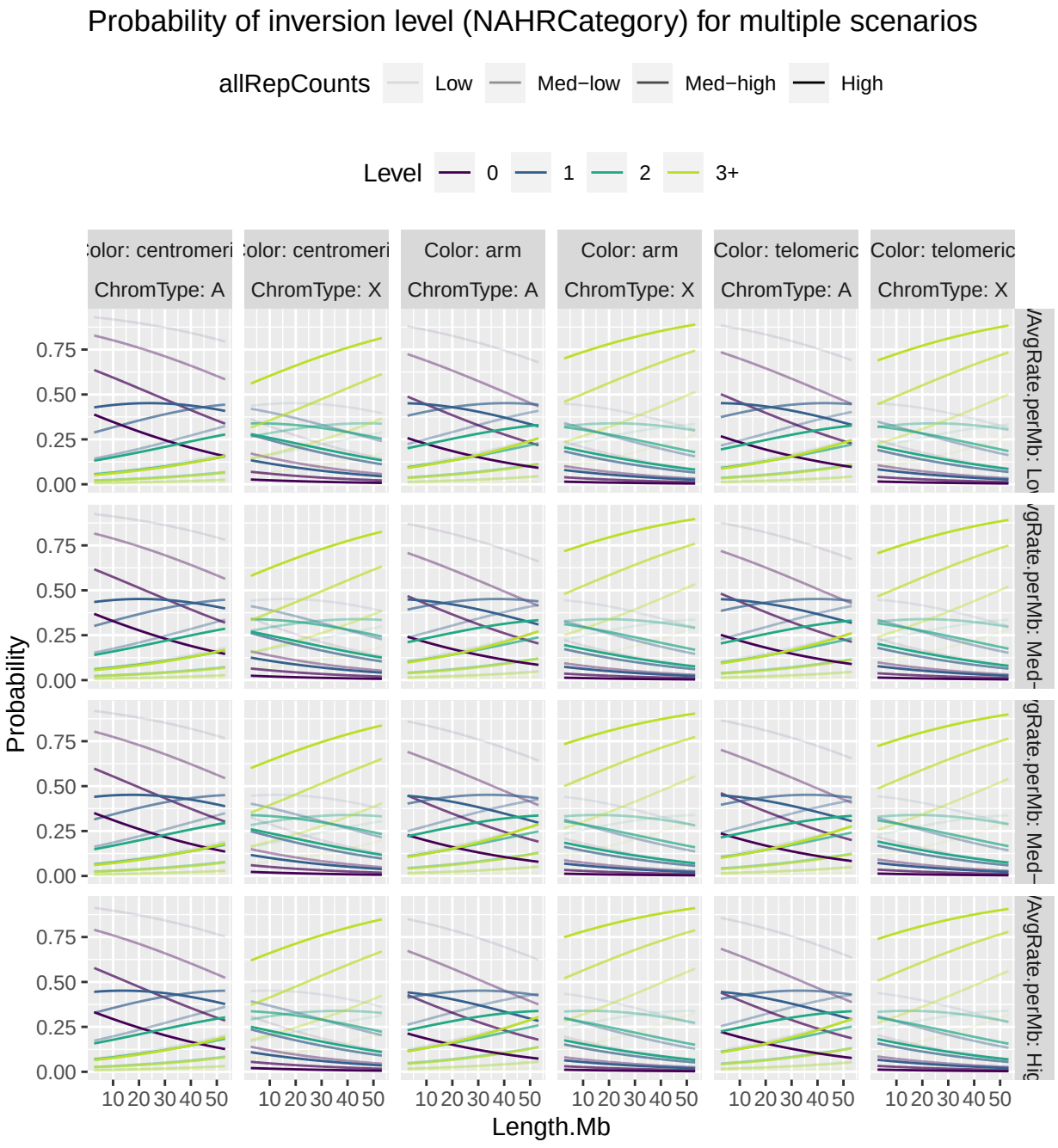


Figure 8: Probabiilty of having 0 to >3 inversions depending on multiple independent variables

Scaled variables

Total inversions model

Model fitting

```
## Call:
## polr(formula = myFormula, data = winRegions, Hess = T)
##
## Coefficients:
##                          Value Std. Error t value
## Length.Mb.Scaled        0.8529     0.2323  3.6708
## allRepCounts.Scaled     0.2201     0.1740  1.2653
## Colorcentromeric        0.4171     0.5639  0.7397
## Colortelomeric          0.1993     0.4608  0.4326
## WAvgRate.perMb.Scaled -0.1095     0.2581 -0.4241
## ChromTypeX              2.0131     0.7767  2.5920
##
## Intercepts:
##      Value   Std. Error t value
## 0|1  -0.0894  0.2606    -0.3431
## 1|2   1.7324  0.3005     5.7658
## 2|3+  2.8511  0.3847     7.4121
##
## Residual Deviance: 309.0743
## AIC: 327.0743
```

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	0.8528766	0.2323435	3.6707572	0.0002418
allRepCounts.Scaled	0.2201369	0.1739851	1.2652634	0.2057769
Colorcentromeric	0.4171384	0.5639014	0.7397365	0.4594599
Colortelomeric	0.1993078	0.4607575	0.4325656	0.6653304
WAvgRate.perMb.Scaled	-0.1094748	0.2581248	-0.4241156	0.6714815
ChromTypeX	2.0131421	0.7766734	2.5920061	0.0095418
0|1	-0.0894114	0.2606313	-0.3430570	0.7315556
1|2	1.7324085	0.3004607	5.7658405	0.0000000
2|3+	2.8511370	0.3846620	7.4120577	0.0000000

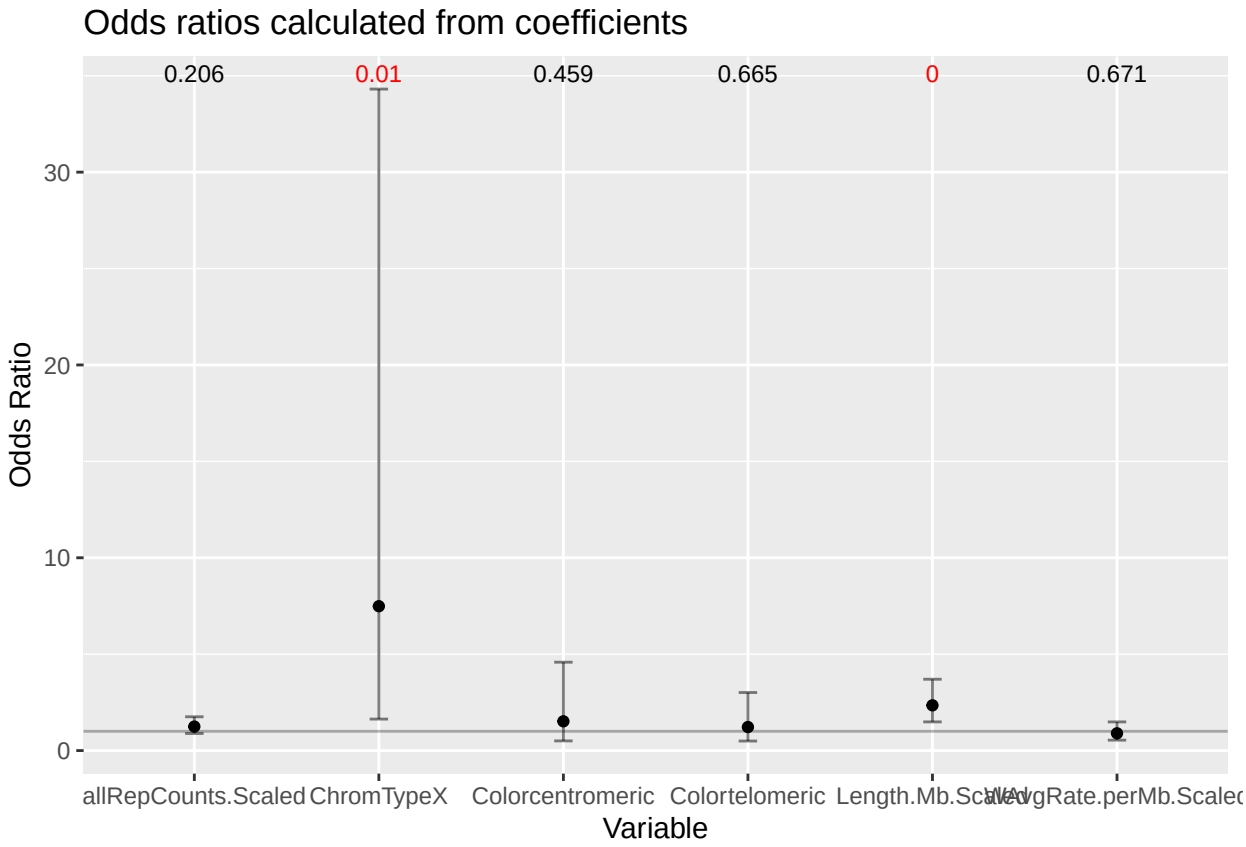
We can also get confidence intervals for the parameter estimates. These can be obtained either by profiling the likelihood function or by using the standard errors and assuming a normal distribution. Note that profiled CIs are not symmetric (although they are usually close to symmetric). If the 95% CI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

	2.5 %	97.5 %
Length.Mb.Scaled	0.3974917	1.3082615
allRepCounts.Scaled	-0.1208675	0.5611414
Colorcentromeric	-0.6880880	1.5223648
Colortelomeric	-0.7037602	1.1023758
WAvgRate.perMb.Scaled	-0.6153901	0.3964406
ChromTypeX	0.4908903	3.5353940

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	2.3463867	1.4880874	3.699736
allRepCounts.Scaled	1.2462474	0.8861513	1.752672
Colorcentromeric	1.5176126	0.5025360	4.583051
Colortelomeric	1.2205576	0.4947216	3.011312
WAvgRate.perMb.Scaled	0.8963048	0.5404300	1.486524
ChromTypeX	7.4868048	1.6337701	34.308528

Example of interpretation: "For 1 unit increase in Length.Mb.Scaled, a window is 2.3463867 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.

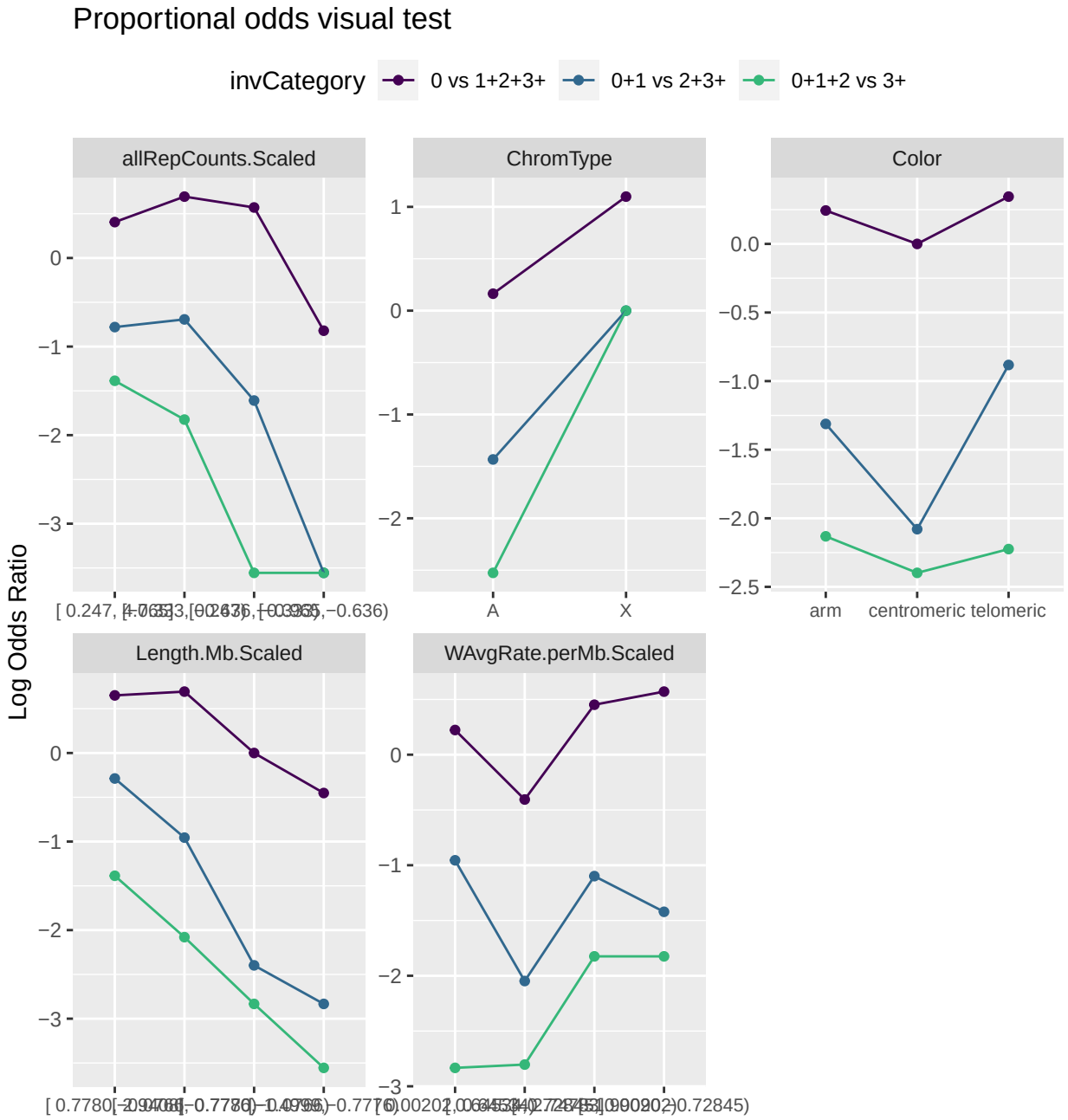
We test the parallel regression assumption with a Brant test:

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

```
## ------------------------------------------------------
## Test for     X2  df  probability
## ------------------------------------------------------
## Omnibus          62.99   12  0
## Length.Mb.Scaled 9.72    2   0.01
## allRepCounts.Scaled  0   2   1
## Colorcentromeric 2.1 2   0.35
## Colortelomeric       1.06    2   0.59
## WAvgRate.perMb.Scaled    0.67    2   0.71
## ChromTypeX       9.99    2   0.01
## ------------------------------------------------------
##
## H0: Parallel Regression Assumption holds
```

	X2	df	probability
Omnibus	62.9898415	12	0.0000000
Length.Mb.Scaled	9.7168082	2	0.0077629
allRepCounts.Scaled	0.0029805	2	0.9985108
Colorcentromeric	2.0999201	2	0.3499517
Colortelomeric	1.0607528	2	0.5883834
WAvgRate.perMb.Scaled	0.6717577	2	0.7147097
ChromTypeX	9.9910725	2	0.0067681

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.

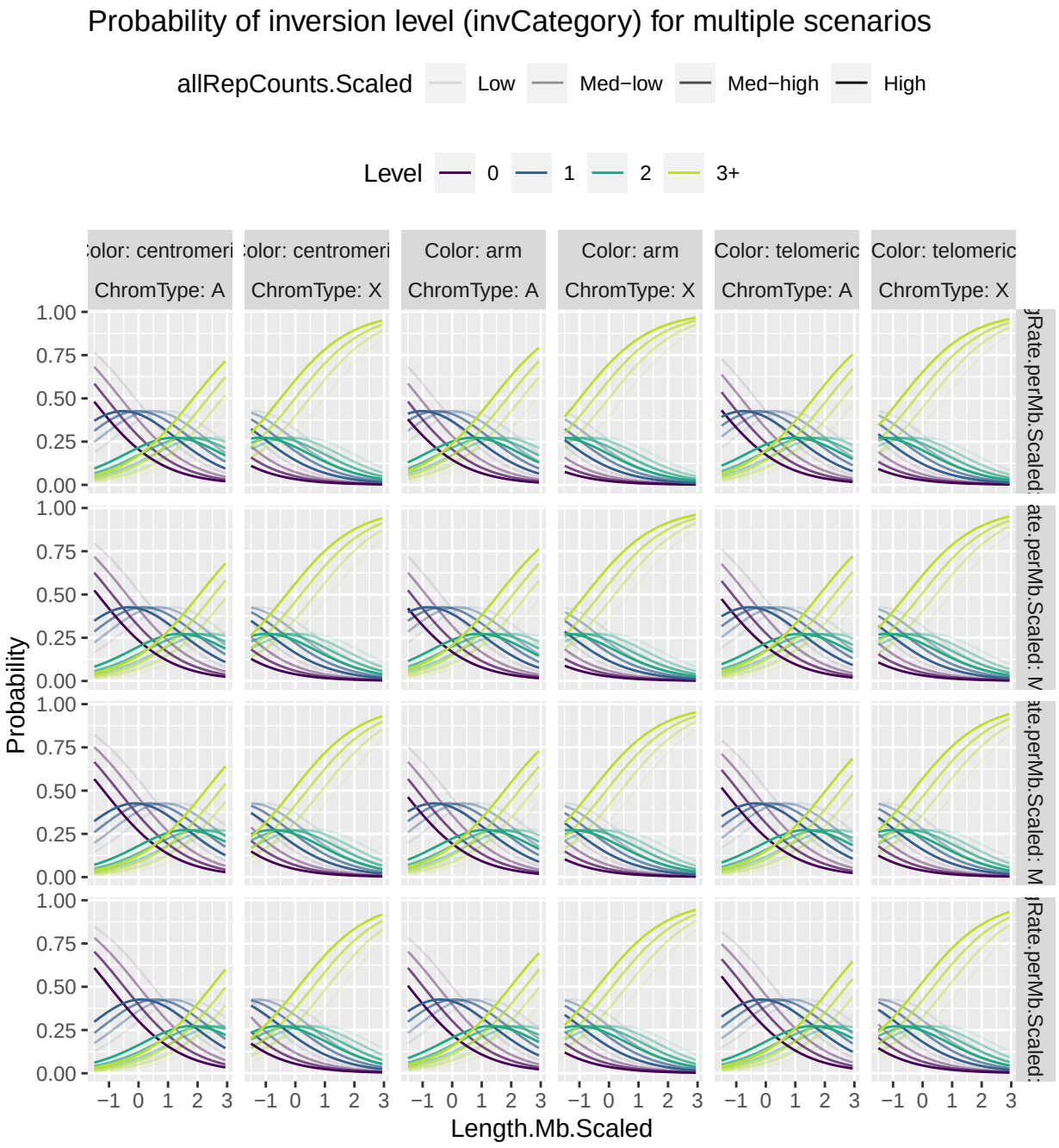


Figure 9: Probabiilty of having 0 to >3 inversions depending on multiple independent variables

NH inversions model

Model fitting

```
## Call:
## polr(formula = myFormula, data = winRegions, Hess = T)
##
## Coefficients:
##                          Value Std. Error t value
## Length.Mb.Scaled        1.0058     0.2580  3.8981
## allRepCounts.Scaled    -0.1960     0.1940 -1.0104
## Colorcentromeric        0.2289     0.6137  0.3729
## Colortelomeric         -0.1214     0.5275 -0.2302
## WAvgRate.perMb.Scaled -0.1752     0.3089 -0.5672
## ChromTypeX             -0.8589     0.8689 -0.9885
##
## Intercepts:
##      Value   Std. Error t value
## 0|1   0.4859  0.2801     1.7344
## 1|2   2.3381  0.3576     6.5380
## 2|3+  3.7289  0.5275     7.0685
##
## Residual Deviance: 248.1198
## AIC: 266.1198
```

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	1.0057585	0.2580094	3.8981475	0.0000969
allRepCounts.Scaled	-0.1959899	0.1939788	-1.0103676	0.3123192
Colorcentromeric	0.2288664	0.6136872	0.3729365	0.7091957
Colortelomeric	-0.1214254	0.5274576	-0.2302088	0.8179295
WAvgRate.perMb.Scaled	-0.1752176	0.3089161	-0.5672012	0.5705775
ChromTypeX	-0.8588814	0.8688626	-0.9885123	0.3229018
0|1	0.4858567	0.2801295	1.7344005	0.0828471
1|2	2.3381314	0.3576191	6.5380494	0.0000000
2|3+	3.7288537	0.5275330	7.0684746	0.0000000

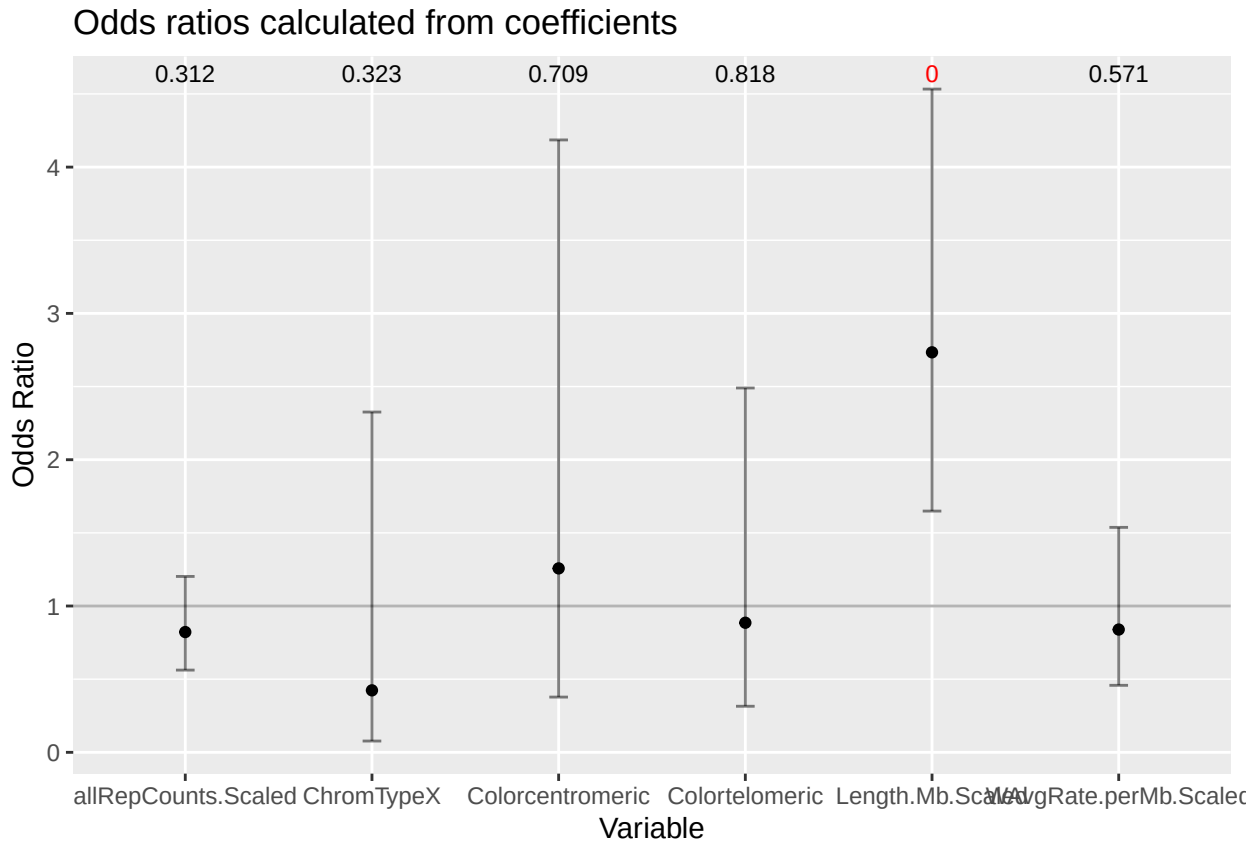
We can also get confidence intervals for the parameter estimates. These can be obtained either by profiling the likelihood function or by using the standard errors and assuming a normal distribution. Note that profiled CIs are not symmetric (although they are usually close to symmetric). If the 95% CI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

	2.5 %	97.5 %
Length.Mb.Scaled	0.5000695	1.5114476
allRepCounts.Scaled	-0.5761812	0.1842015
Colorcentromeric	-0.9739384	1.4316711
Colortelomeric	-1.1552233	0.9123725
WAvgRate.perMb.Scaled	-0.7806821	0.4302469
ChromTypeX	-2.5618207	0.8440580

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	2.7339803	1.6488358	4.533288
allRepCounts.Scaled	0.8220206	0.5620406	1.202258
Colorcentromeric	1.2571740	0.3775930	4.185688
Colortelomeric	0.8856571	0.3149872	2.490224
WAvgRate.perMb.Scaled	0.8392744	0.4580934	1.537637
ChromTypeX	0.4236357	0.0771641	2.325786

Example of interpretation: "For 1 unit increase in Length.Mb.Scaled, a window is 2.7339803 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.

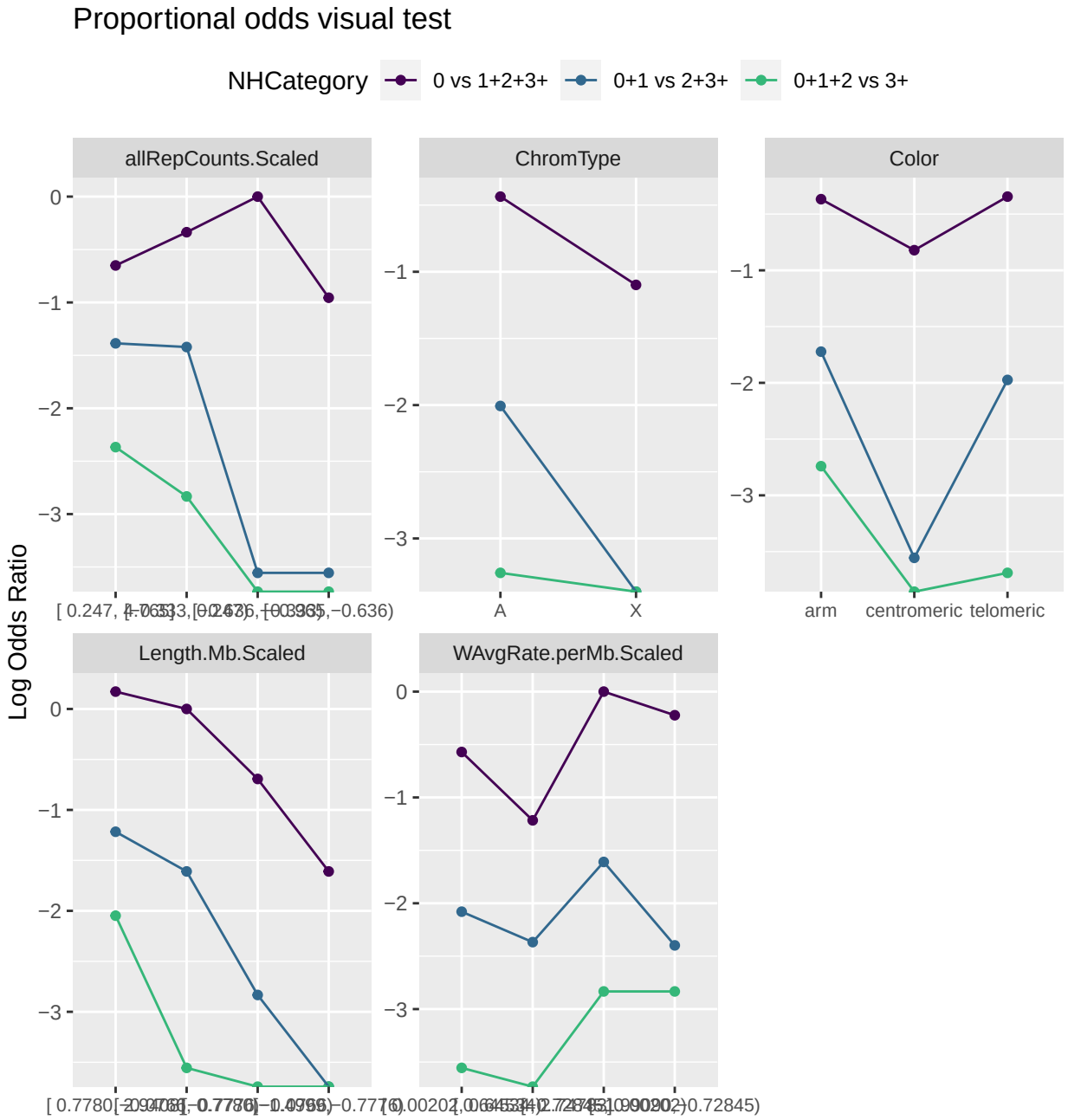
We test the parallel regression assumption with a Brant test:

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

```
## ------------------------------------------------------
## Test for     X2  df  probability
## ------------------------------------------------------
## Omnibus          19.16   12  0.08
## Length.Mb.Scaled 3.58    2   0.17
## allRepCounts.Scaled  1.09    2   0.58
## Colorcentromeric 0.12    2   0.94
## Colortelomeric       4.89    2   0.09
## WAvgRate.perMb.Scaled    8.45    2   0.01
## ChromTypeX       0   2   1
## ------------------------------------------------------
##
## H0: Parallel Regression Assumption holds
```

	X2	df	probability
Omnibus	19.1612073	12	0.0847116
Length.Mb.Scaled	3.5779305	2	0.1671330
allRepCounts.Scaled	1.0895693	2	0.5799667
Colorcentromeric	0.1183776	2	0.9425288
Colortelomeric	4.8901861	2	0.0867181
WAvgRate.perMb.Scaled	8.4547485	2	0.0145907
ChromTypeX	0.0001120	2	0.9999440

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.

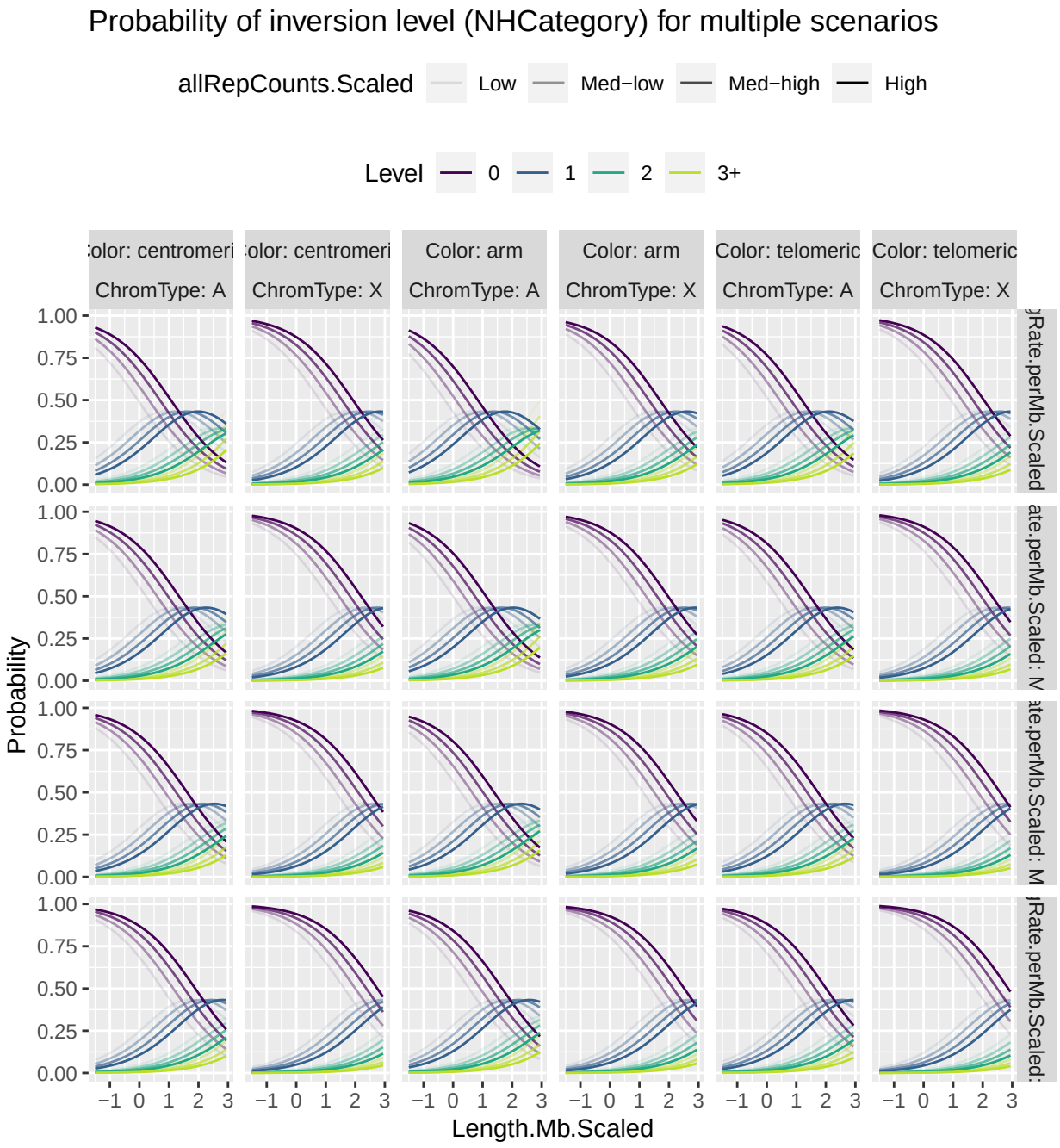


Figure 10: Probabiilty of having 0 to >3 inversions depending on multiple independent variables

NAHR inversions model

Model fitting

```
## Call:
## polr(formula = myFormula, data = winRegions, Hess = T)
##
## Coefficients:
##                          Value Std. Error t value
## Length.Mb.Scaled      0.27779     0.2691  1.0321
## allRepCounts.Scaled   0.53082     0.2011  2.6389
## Colorcentromeric      0.60566     0.7235  0.8371
## Colortelomeric        0.55162     0.5581  0.9883
## WAvgRate.perMb.Scaled 0.05175     0.3317  0.1560
## ChromTypeX            3.14870     0.8529  3.6919
##
## Intercepts:
##      Value  Std. Error t value
## 0|1  1.5546 0.3317     4.6865
## 1|2  3.5017 0.4857     7.2099
## 2|3+ 4.9147 0.7398     6.6431
##
## Residual Deviance: 194.8853
## AIC: 212.8853
```

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	0.2777872	0.2691485	1.0320963	0.3020270
allRepCounts.Scaled	0.5308151	0.2011476	2.6389333	0.0083167
Colorcentromeric	0.6056551	0.7235398	0.8370723	0.4025519
Colortelomeric	0.5516164	0.5581217	0.9883443	0.3229840
WAvgRate.perMb.Scaled	0.0517493	0.3316889	0.1560176	0.8760192
ChromTypeX	3.1486998	0.8528775	3.6918548	0.0002226
0|1	1.5545529	0.3317052	4.6865495	0.0000028
1|2	3.5016706	0.4856745	7.2099122	0.0000000
2|3+	4.9147482	0.7398288	6.6430888	0.0000000

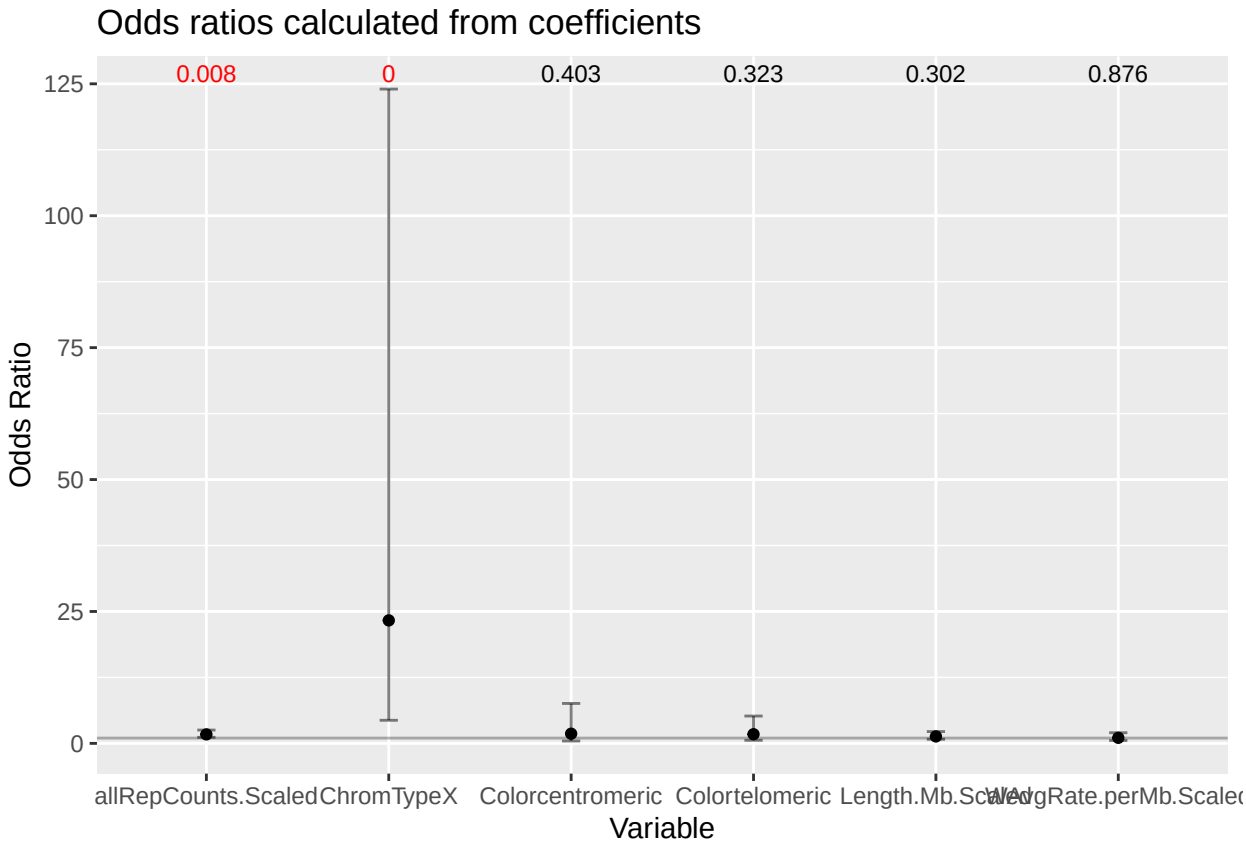
We can also get confidence intervals for the parameter estimates. These can be obtained either by profiling the likelihood function or by using the standard errors and assuming a normal distribution. Note that profiled CIs are not symmetric (although they are usually close to symmetric). If the 95% CI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

	2.5 %	97.5 %
Length.Mb.Scaled	-0.2497342	0.8053085
allRepCounts.Scaled	0.1365730	0.9250571
Colorcentromeric	-0.8124568	2.0237670
Colortelomeric	-0.5422820	1.6455149
WAvgRate.perMb.Scaled	-0.5983489	0.7018475
ChromTypeX	1.4770907	4.8203089

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	1.320205	0.7790078	2.237387
allRepCounts.Scaled	1.700318	1.1463386	2.522012
Colorcentromeric	1.832452	0.4437665	7.566775
Colortelomeric	1.736057	0.5814199	5.183678
WAvgRate.perMb.Scaled	1.053112	0.5497185	2.017477
ChromTypeX	23.305742	4.3801837	124.003384

Example of interpretation: "For 1 unit increase in Length.Mb.Scaled, a window is 1.3202052 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.

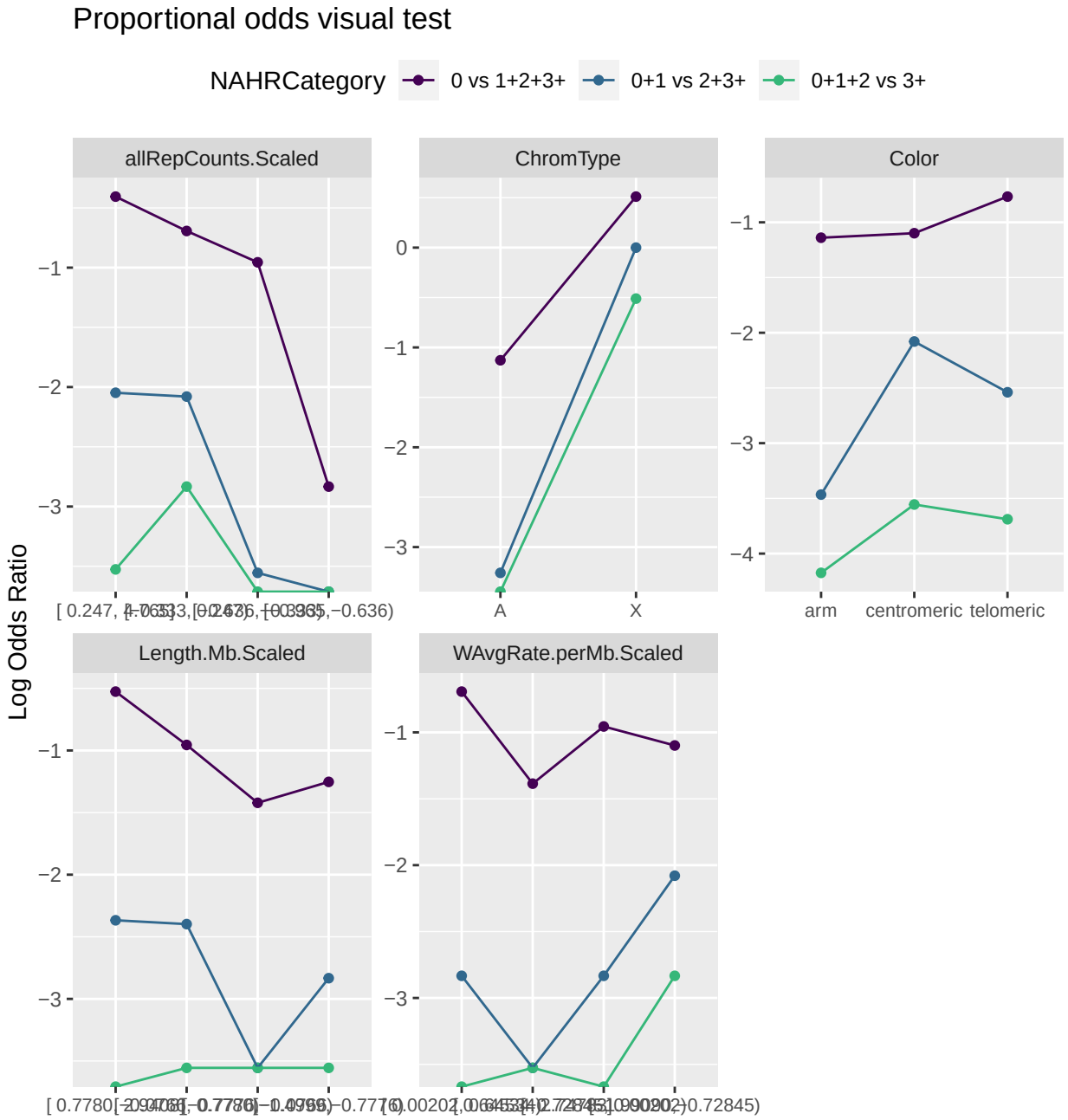
We test the parallel regression assumption with a Brant test:

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

```
## ------------------------------------------------------
## Test for     X2  df  probability
## ------------------------------------------------------
## Omnibus          17.32   12  0.14
## Length.Mb.Scaled 0   2   1
## allRepCounts.Scaled  0   2   1
## Colorcentromeric 0   2   1
## Colortelomeric       0   2   1
## WAvgRate.perMb.Scaled    0   2   1
## ChromTypeX       0   2   1
## ------------------------------------------------------
##
## H0: Parallel Regression Assumption holds
```

	X2	df	probability
Omnibus	17.3245197	12	0.1377923
Length.Mb.Scaled	0.0000001	2	0.9999999
allRepCounts.Scaled	-0.0000060	2	1.0000000
Colorcentromeric	0.0000000	2	1.0000000
Colortelomeric	-0.0000001	2	1.0000000
WAvgRate.perMb.Scaled	0.0000001	2	0.9999999
ChromTypeX	-0.0000050	2	1.0000000

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.

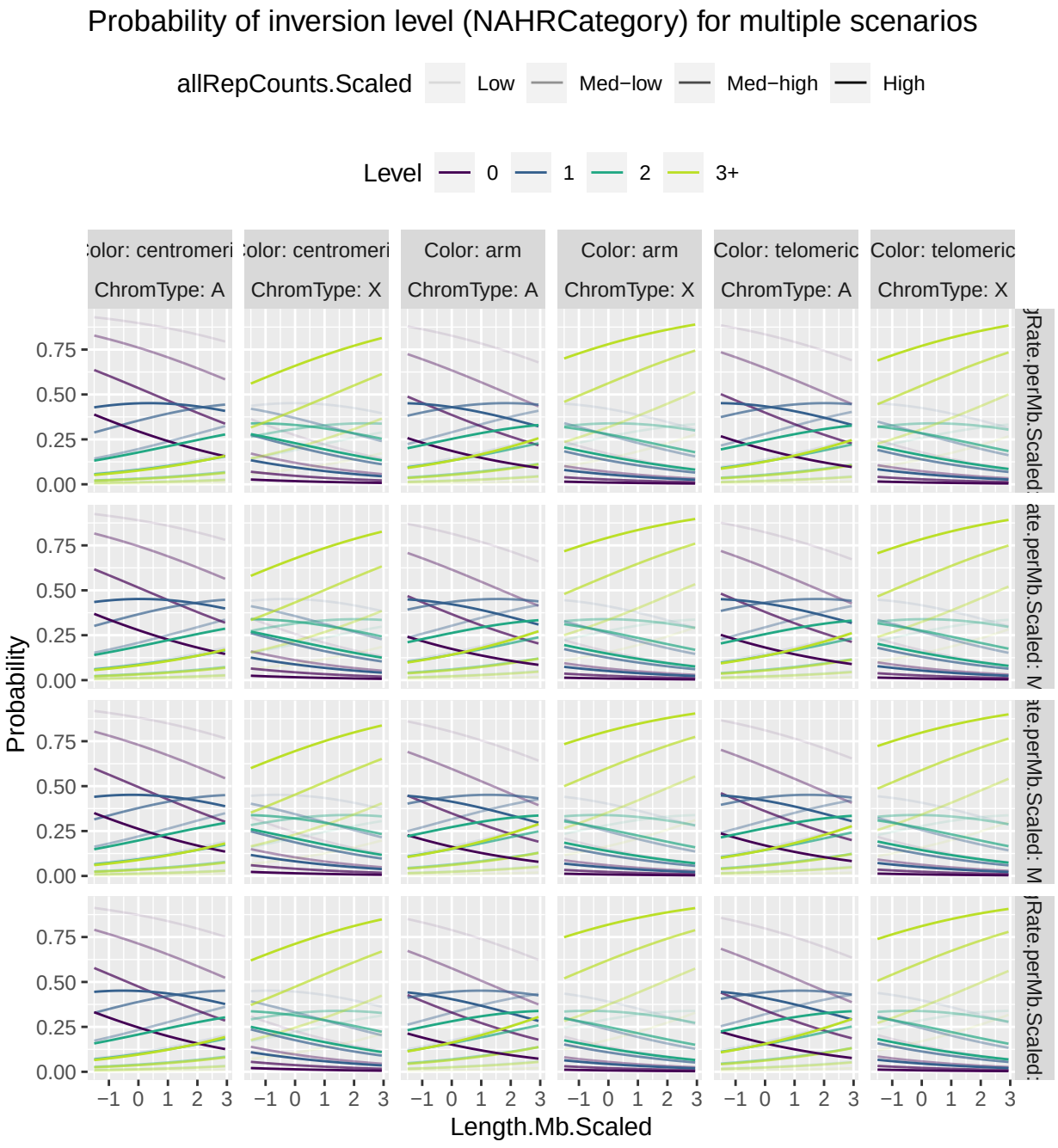


Figure 11: Probabiilty of having 0 to >3 inversions depending on multiple independent variables

Descriptive categories

Descriptive statistics

Raw data:

Chromosome	Start	End	Color	invCenter	NHCenter	NAHRCenter	Length.Mb	RepCounts	log10RepCounts	WAvgRate.perMb	ChrType
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	113381273	135473442	telomeric	1	1	0	22.092163	170	2.230449	2.1846155	A
chr10	42436305	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.

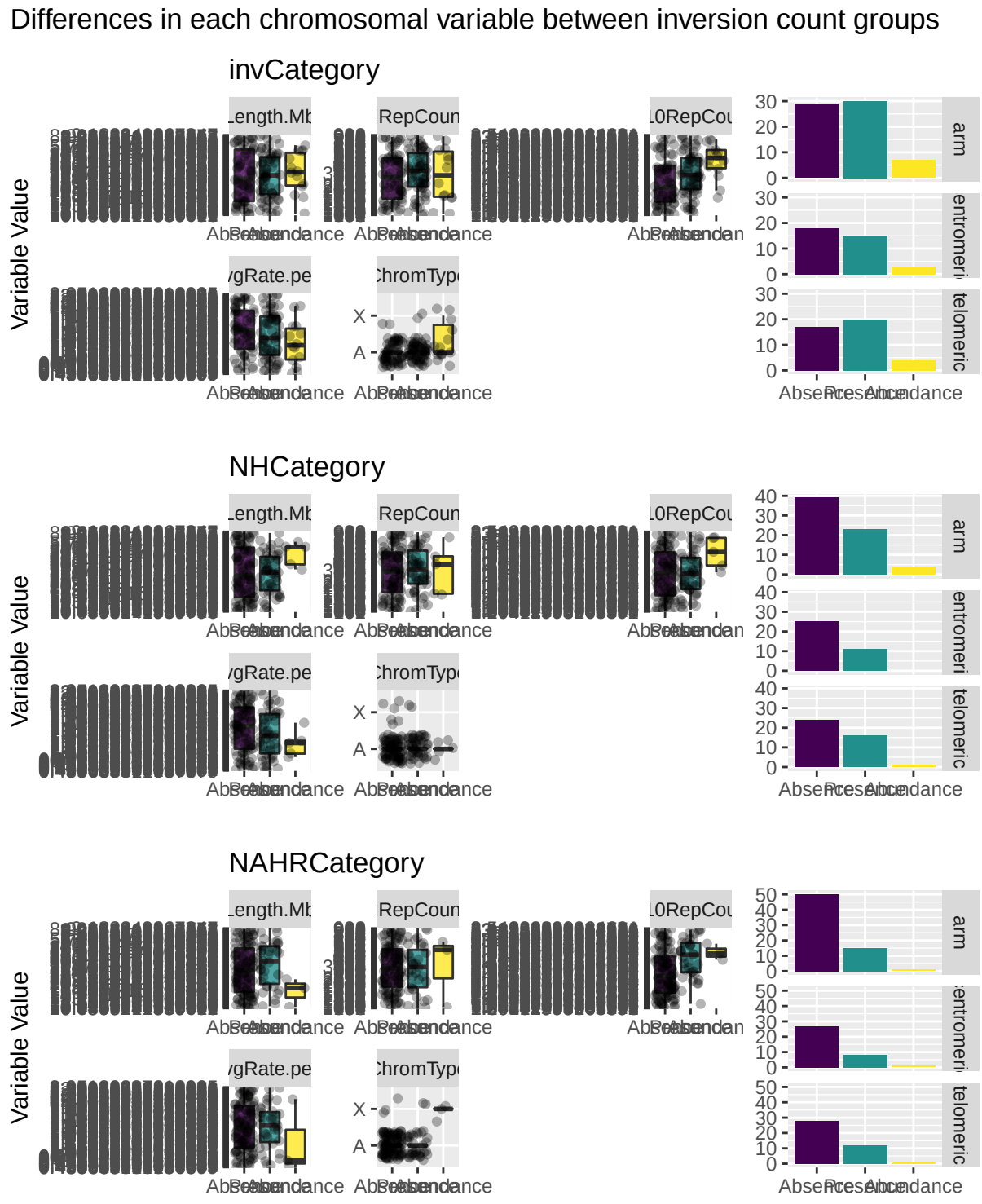


Figure 13: Potential effect of independent variables on the different types of invesions.

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

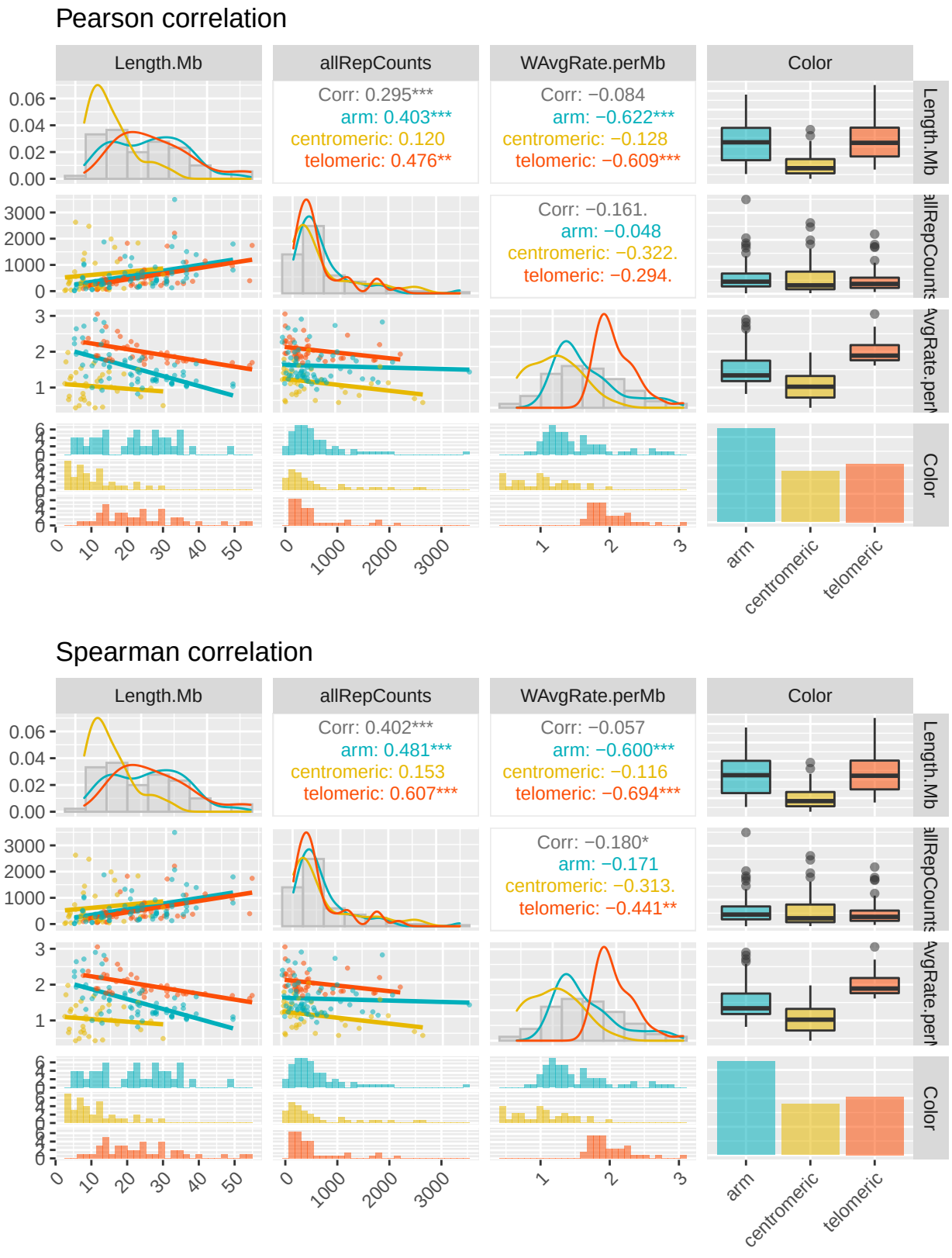


Figure 14: Correlations between variables.

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         0.0648463  0.0213170  3.0420
## allRepCounts      0.0003299  0.0003144  1.0492
## Colorcentromeric  0.3582519  0.5869634  0.6103
## Colortelomeric    0.1060988  0.4751575  0.2233
## WAvgRate.perMb   -0.2772739  0.4824877 -0.5747
## ChromTypeX        2.2374275  0.8186743  2.7330
##
## Intercepts:
##                     Value   Std. Error t value
## Absence|Presence     0.9720  1.0738     0.9052
## Presence|Abundance   3.8245  1.1441     3.3429
##
## Residual Deviance: 242.4534
## AIC: 258.4534
```

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

	Value	Std. Error	t value	p value
Length.Mb	0.0648463	0.0213170	3.0419990	0.0023501
allRepCounts	0.0003299	0.0003144	1.0492492	0.2940634
Colorcentromeric	0.3582519	0.5869634	0.6103479	0.5416314
Colortelomeric	0.1060988	0.4751575	0.2232918	0.8233084
WAvgRate.perMb	-0.2772739	0.4824877	-0.5746756	0.5655107
ChromTypeX	2.2374275	0.8186743	2.7329885	0.0062762
Absence|Presence	0.9719916	1.0738237	0.9051687	0.3653760
Presence|Abundance	3.8245346	1.1440601	3.3429489	0.0008289

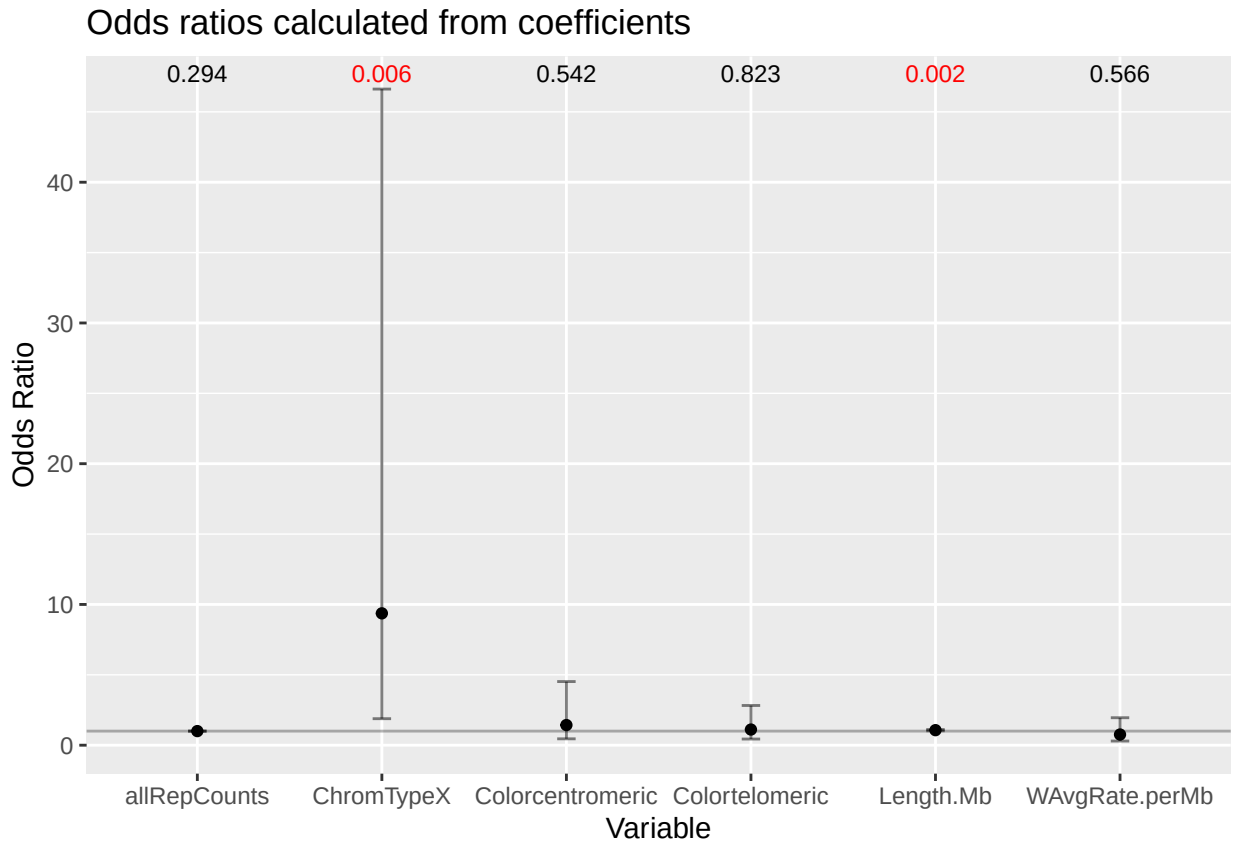
We can also get confidence intervals for the parameter estimates. These can be obtained either by profiling the likelihood function or by using the standard errors and assuming a normal distribution. Note that profiled CIs are not symmetric (although they are usually close to symmetric). If the 95% CI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

	2.5 %	97.5 %
Length.Mb	0.0230658	0.1066269
allRepCounts	-0.0002863	0.0009461
Colorcentromeric	-0.7921753	1.5086791
Colortelomeric	-0.8251928	1.0373904
WAvgRate.perMb	-1.2229323	0.6683846
ChromTypeX	0.6328553	3.8419997

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	1.0669951	1.0233338	1.112519
allRepCounts	1.0003299	0.9997137	1.000946
Colorcentromeric	1.4308260	0.4528586	4.520755
Colortelomeric	1.1119317	0.4381505	2.821843
WAvgRate.perMb	0.7578469	0.2943657	1.951083
ChromTypeX	9.3691981	1.8829794	46.618605

Example of interpretation: "For 1 unit increase in Length.Mb, a window is 1.0669951 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.

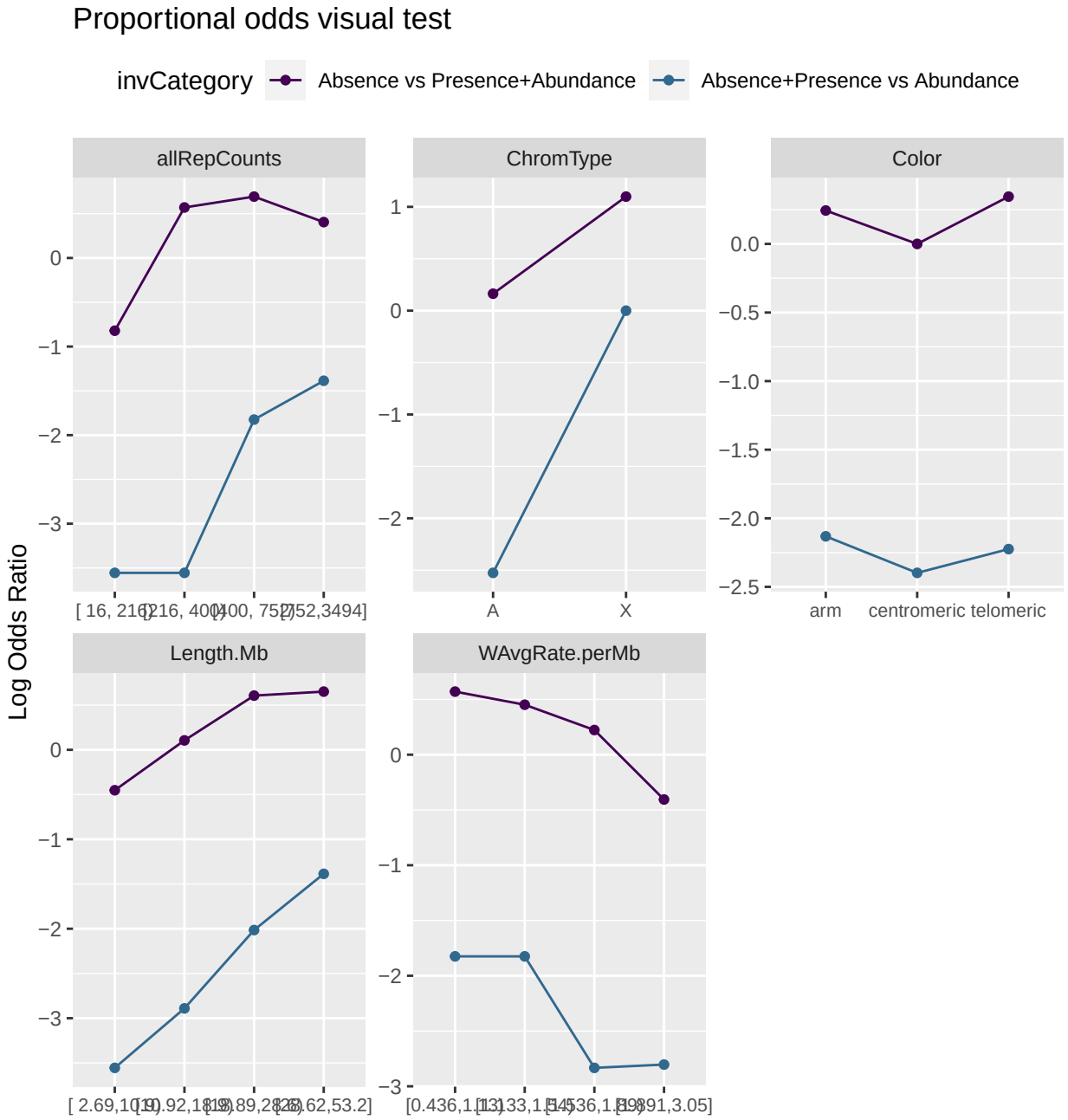
We test the parallel regression assumption with a Brant test:

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

```
## ------------------------------------------------------
## Test for     X2  df  probability
## ------------------------------------------------------
## Omnibus          10.79   6   0.09
## Length.Mb        6.05    1   0.01
## allRepCounts     0   1   0.96
## Colorcentromeric 2.1 1   0.15
## Colortelomeric       0.71    1   0.4
## WAvgRate.perMb       0.42    1   0.52
## ChromTypeX       8.7 1   0
## ------------------------------------------------------
##
## H0: Parallel Regression Assumption holds
```

	X2	df	probability
Omnibus	10.7937767	6	0.0949630
Length.Mb	6.0511470	1	0.0138973
allRepCounts	0.0029780	1	0.9564800
Colorcentromeric	2.0995048	1	0.1473469
Colortelomeric	0.7113203	1	0.3990059
WAvgRate.perMb	0.4197315	1	0.5170711
ChromTypeX	8.7025080	1	0.0031777

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.

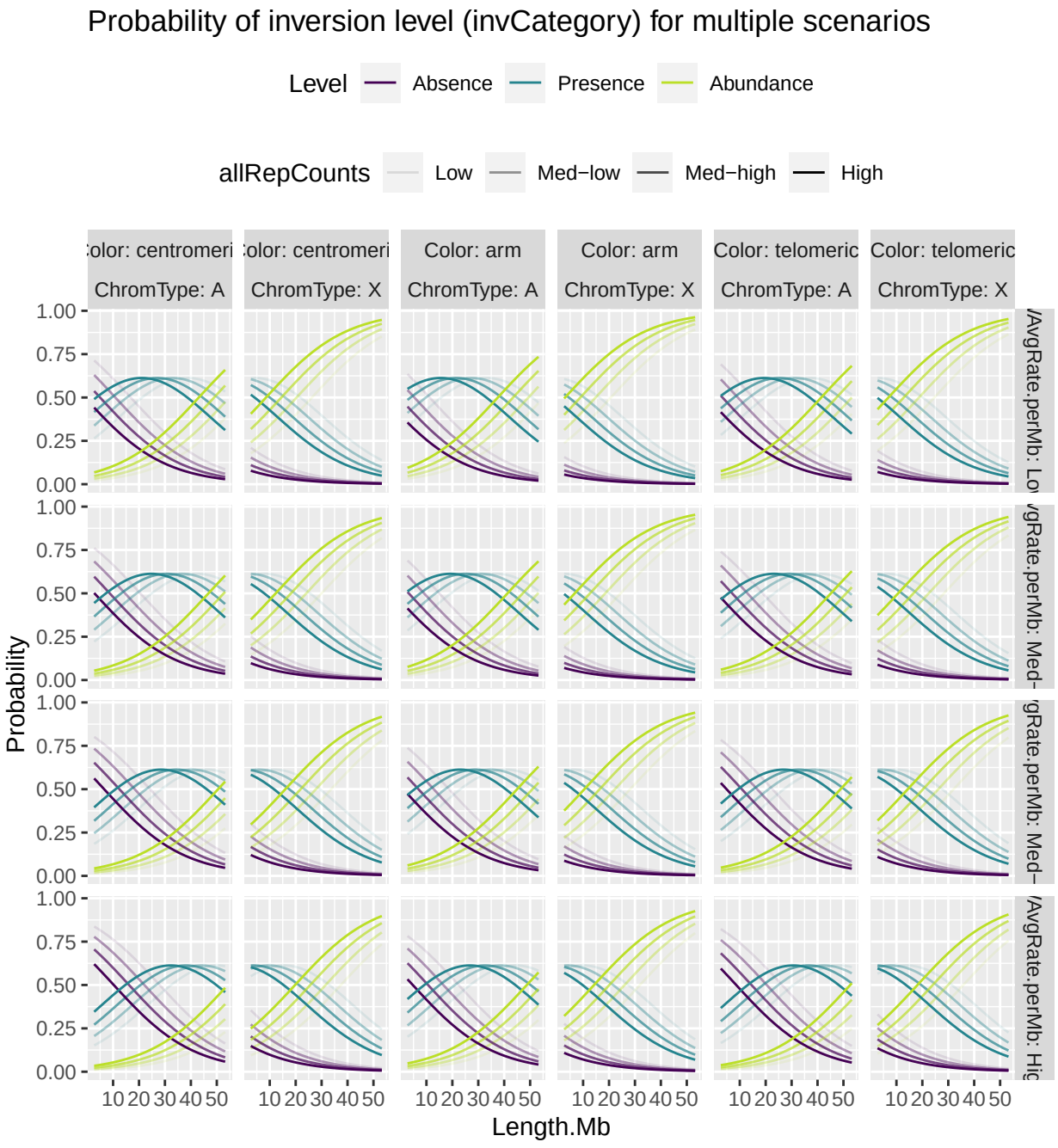


Figure 15: Probabiilty of having 0 to >3 inversions depending on multiple independent variables

NH inversions model

Model fitting

```
## Call:
## polr(formula = myFormula, data = winRegions, Hess = T)
##
## Coefficients:
##                       Value Std. Error t value
## Length.Mb         0.0881632  0.0237182  3.7171
## allRepCounts     -0.0004447  0.0003647 -1.2195
## Colorcentromeric  0.1849909  0.6328854  0.2923
## Colortelomeric    0.0221750  0.5422100  0.0409
## WAvgRate.perMb   -0.5599917  0.5803706 -0.9649
## ChromTypeX       -0.8531096  0.8893640 -0.9592
##
## Intercepts:
##                     Value   Std. Error t value
## Absence|Presence    1.1405  1.2078     0.9443
## Presence|Abundance  4.4203  1.3328     3.3164
##
## Residual Deviance: 195.6799
## AIC: 211.6799
```

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

	Value	Std. Error	t value	p value
Length.Mb	0.0881632	0.0237182	3.7171152	0.0002015
allRepCounts	-0.0004447	0.0003647	-1.2194830	0.2226609
Colorcentromeric	0.1849909	0.6328854	0.2922977	0.7700590
Colortelomeric	0.0221750	0.5422100	0.0408975	0.9673776
WAvgRate.perMb	-0.5599917	0.5803706	-0.9648863	0.3346018
ChromTypeX	-0.8531096	0.8893640	-0.9592355	0.3374401
Absence|Presence	1.1404784	1.2078002	0.9442608	0.3450364
Presence|Abundance	4.4202568	1.3328323	3.3164388	0.0009117

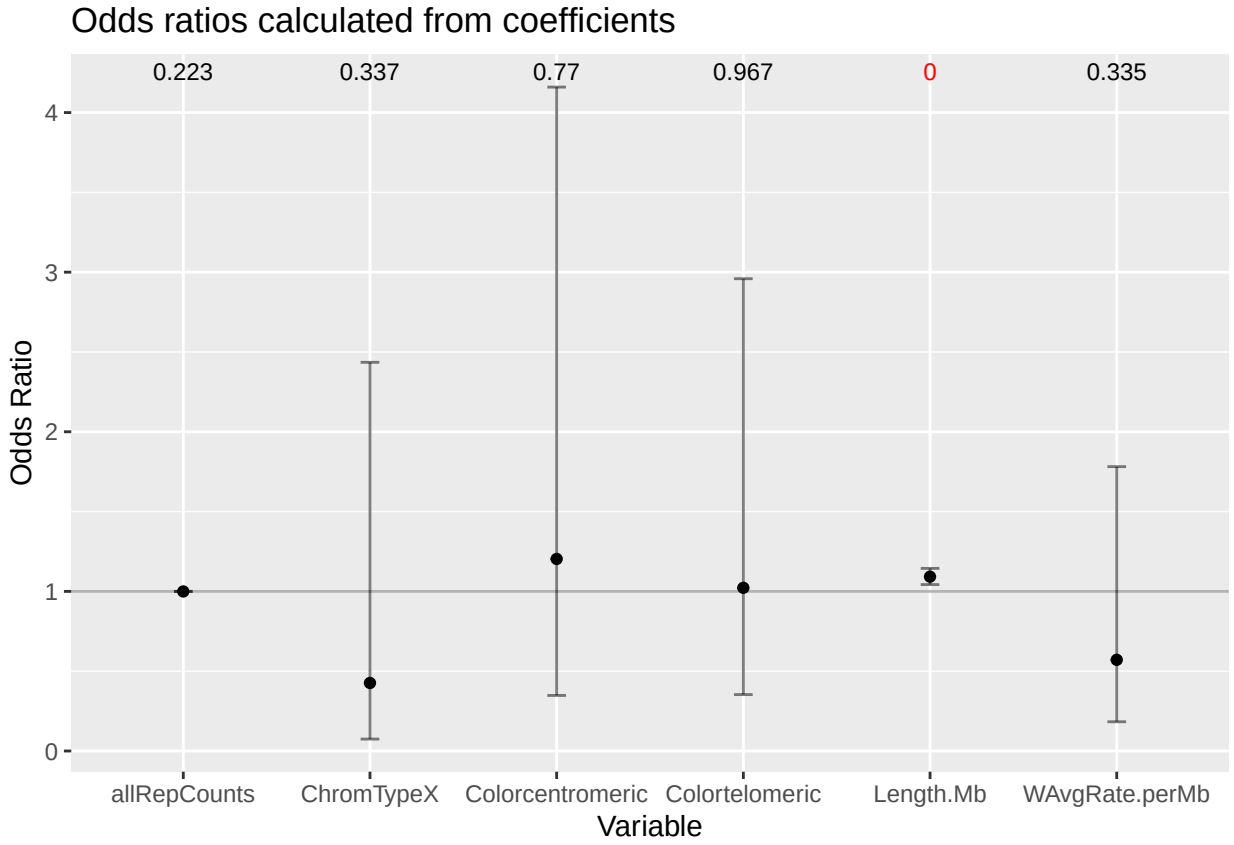
We can also get confidence intervals for the parameter estimates. These can be obtained either by profiling the likelihood function or by using the standard errors and assuming a normal distribution. Note that profiled CIs are not symmetric (although they are usually close to symmetric). If the 95% CI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

	2.5 %	97.5 %
Length.Mb	0.0416764	0.1346500
allRepCounts	-0.0011594	0.0002700
Colorcentromeric	-1.0554416	1.4254235
Colortelomeric	-1.0405370	1.0848871
WAvgRate.perMb	-1.6974972	0.5775139
ChromTypeX	-2.5962311	0.8900119

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	1.0921664	1.0425571	1.144136
allRepCounts	0.9995554	0.9988413	1.000270
Colorcentromeric	1.2032075	0.3480387	4.159619
Colortelomeric	1.0224227	0.3532649	2.959106
WAvgRate.perMb	0.5712138	0.1831413	1.781604
ChromTypeX	0.4260879	0.0745540	2.435159

Example of interpretation: "For 1 unit increase in Length.Mb, a window is 1.0921664 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:

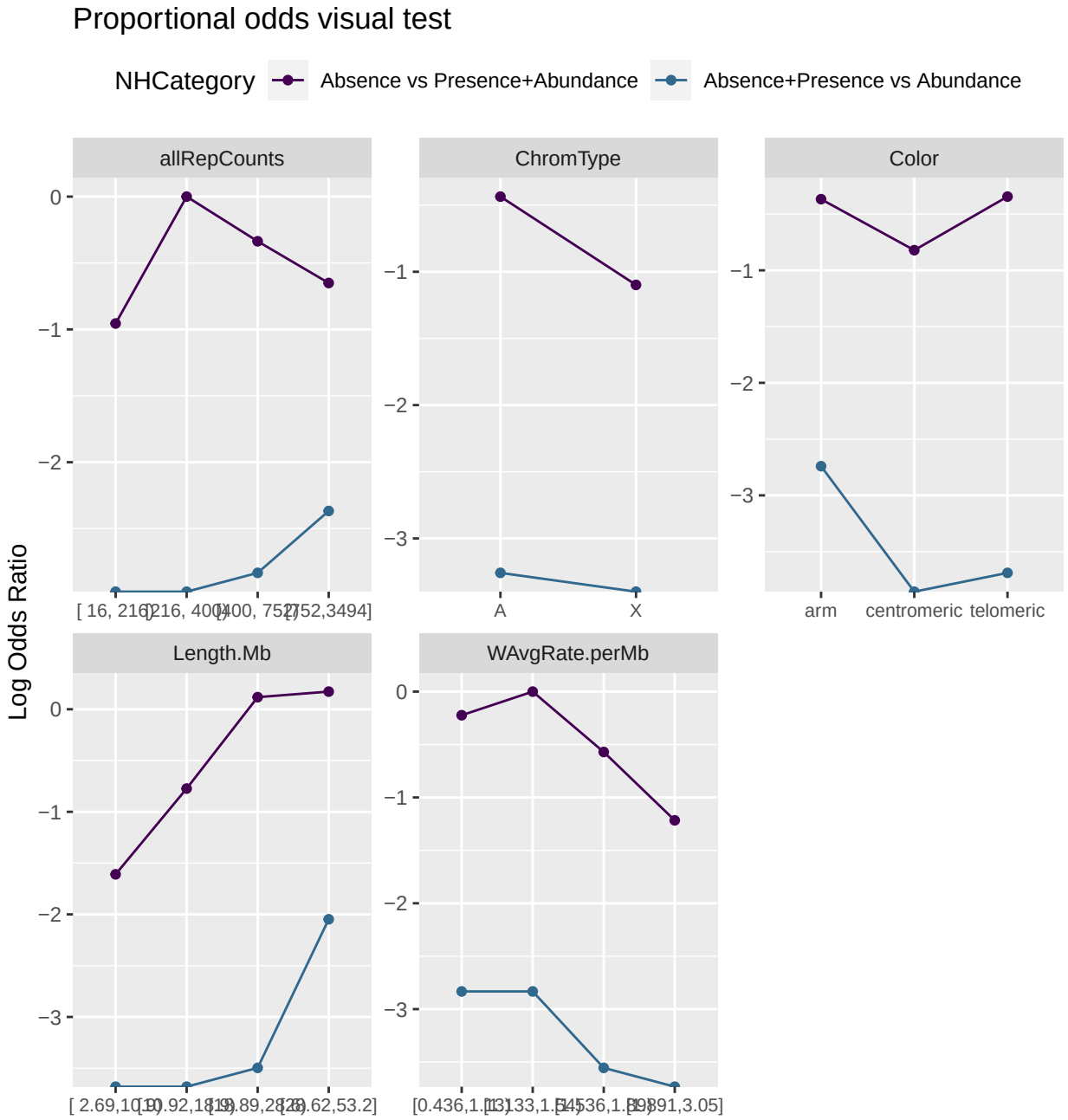
```
pacman::p_load("brant", "Hmisc")
btest<-brant(mod)
```

```
## ------------------------------------------------------
## Test for     X2  df  probability
## ------------------------------------------------------
```

```
## Omnibus          4.34    6   0.63
## Length.Mb        2.34    1   0.13
## allRepCounts     0.37    1   0.55
## Colorcentromeric 0   1   1
## Colortelomeric       0.78    1   0.38
## WAvgRate.perMb       0.01    1   0.93
## ChromTypeX       0   1   1
## ----------------------------------------------------
##
## H0: Parallel Regression Assumption holds
```

	X2	df	probability
Omnibus	4.3357879	6	0.6313347
Length.Mb	2.3371907	1	0.1263172
allRepCounts	0.3662969	1	0.5450297
Colorcentromeric	0.0000132	1	0.9971010
Colortelomeric	0.7810778	1	0.3768117
WAvgRate.perMb	0.0079667	1	0.9288782
ChromTypeX	0.0000033	1	0.9985428

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.

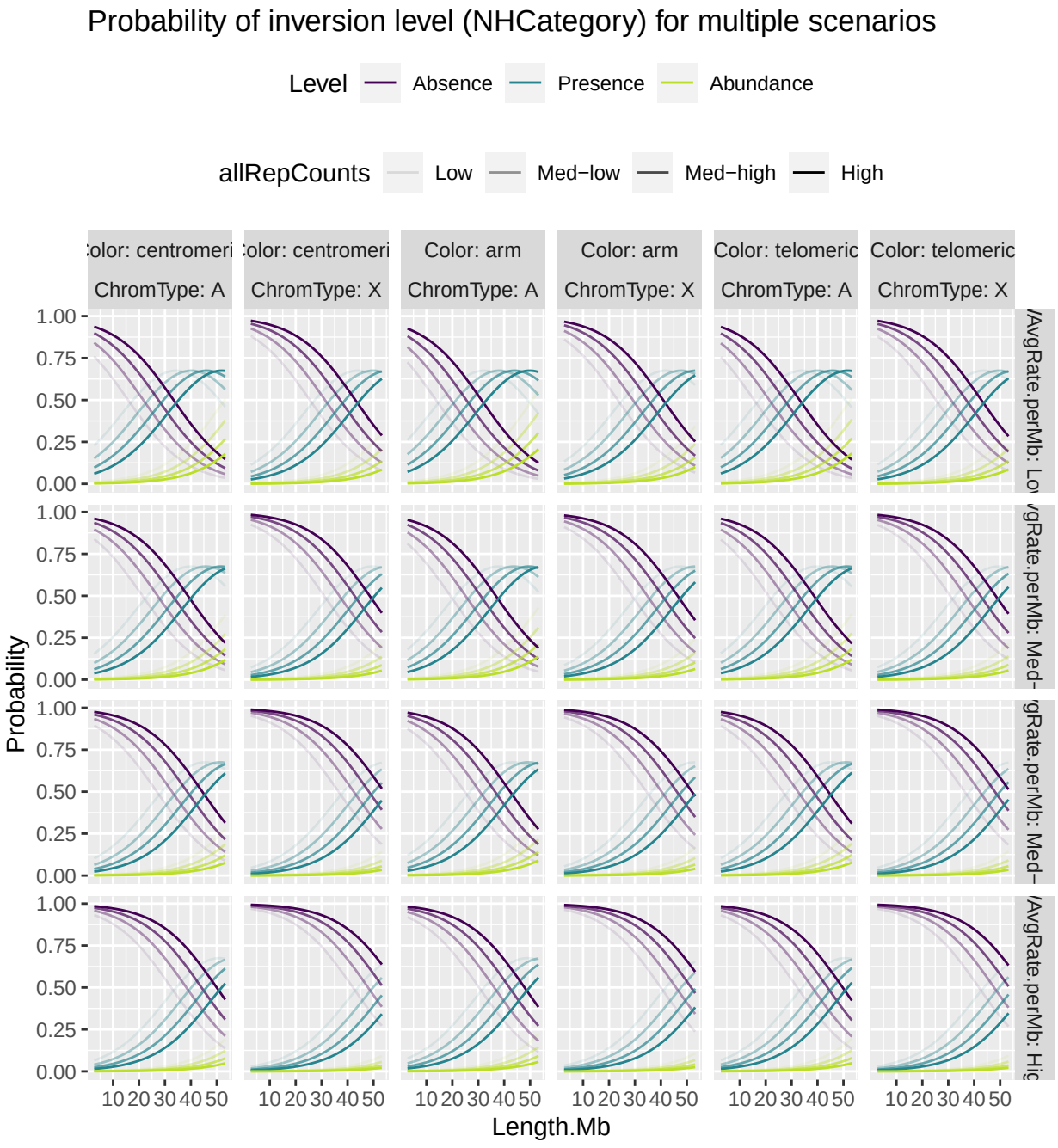


Figure 16: Probabiilty of having 0 to >3 inversions depending on multiple independent variables

NAHR inversions model

Model fitting

```
## Call:
## polr(formula = myFormula, data = winRegions, Hess = T)
##
## Coefficients:
##                      Value Std. Error t value
## Length.Mb        0.0231024   0.024028  0.9615
## allRepCounts     0.0007353   0.000369  1.9928
## Colorcentromeric 0.3953894   0.726258  0.5444
## Colortelomeric   0.4640769   0.555281  0.8358
## WAvgRate.perMb   0.0934910   0.592376  0.1578
## ChromTypeX       3.0023034   0.916542  3.2757
##
## Intercepts:
##                     Value  Std. Error t value
## Absence|Presence   2.5192 1.3241     1.9026
## Presence|Abundance 5.7408 1.4843     3.8676
##
## Residual Deviance: 166.7064
## AIC: 182.7064
```

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

	Value	Std. Error	t value	p value
Length.Mb	0.0231024	0.0240280	0.9614763	0.3363128
allRepCounts	0.0007353	0.0003690	1.9927722	0.0462864
Colorcentromeric	0.3953894	0.7262578	0.5444203	0.5861523
Colortelomeric	0.4640769	0.5552813	0.8357511	0.4032949
WAvgRate.perMb	0.0934910	0.5923755	0.1578239	0.8745956
ChromTypeX	3.0023034	0.9165423	3.2756844	0.0010541
Absence|Presence	2.5192498	1.3241418	1.9025529	0.0570989
Presence|Abundance	5.7408086	1.4843305	3.8676081	0.0001099

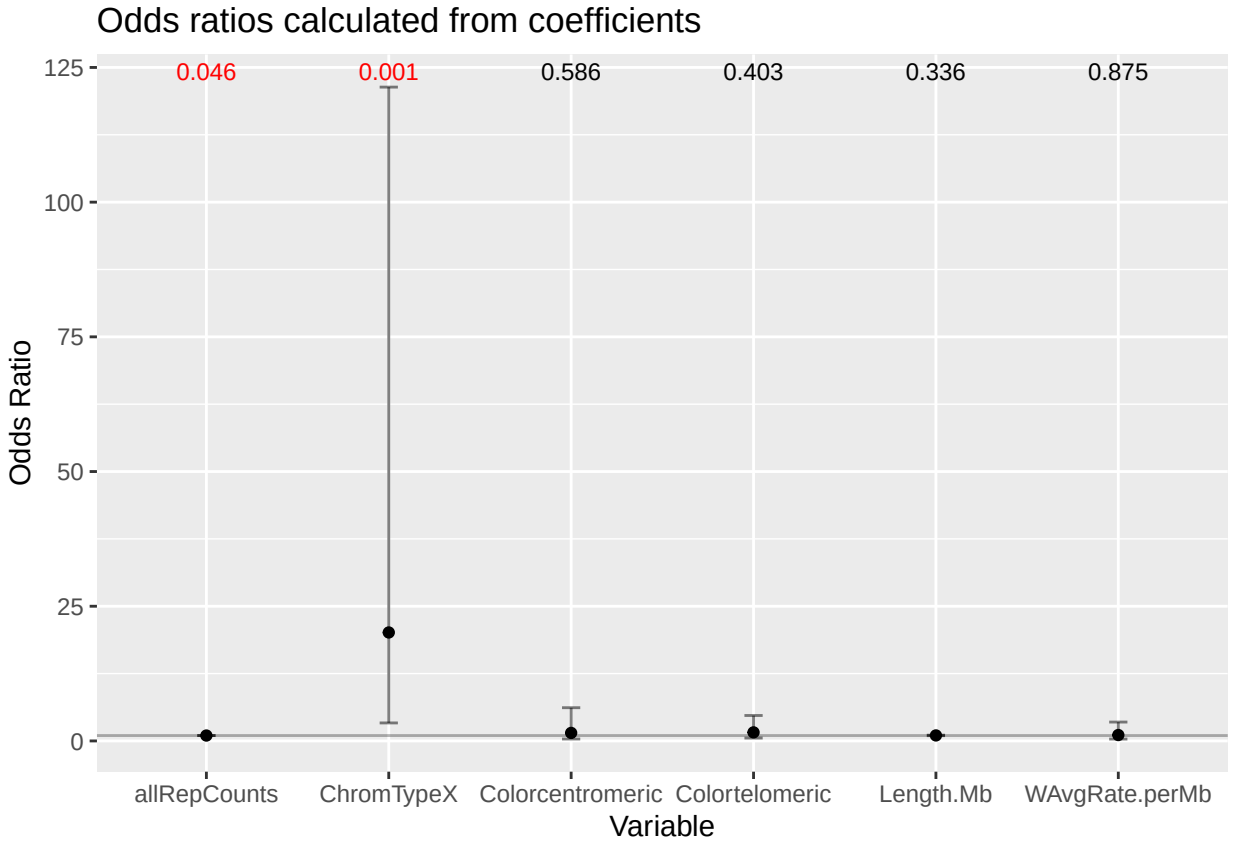
We can also get confidence intervals for the parameter estimates. These can be obtained either by profiling the likelihood function or by using the standard errors and assuming a normal distribution. Note that profiled CIs are not symmetric (although they are usually close to symmetric). If the 95% CI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

	2.5 %	97.5 %
Length.Mb	-0.0239917	0.0701964
allRepCounts	0.0000121	0.0014586
Colorcentromeric	-1.0280496	1.8188285
Colortelomeric	-0.6242543	1.5524082
WAvgRate.perMb	-1.0675437	1.2545257
ChromTypeX	1.2059135	4.7986934

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	1.023371	0.9762938	1.072719
allRepCounts	1.000736	1.0000121	1.001460
Colorcentromeric	1.484962	0.3577039	6.164632
Colortelomeric	1.590545	0.5356607	4.722830
WAvgRate.perMb	1.098001	0.3438521	3.506175
ChromTypeX	20.131856	3.3398085	121.351755

Example of interpretation: "For 1 unit increase in Length.Mb, a window is 1.0233713 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:

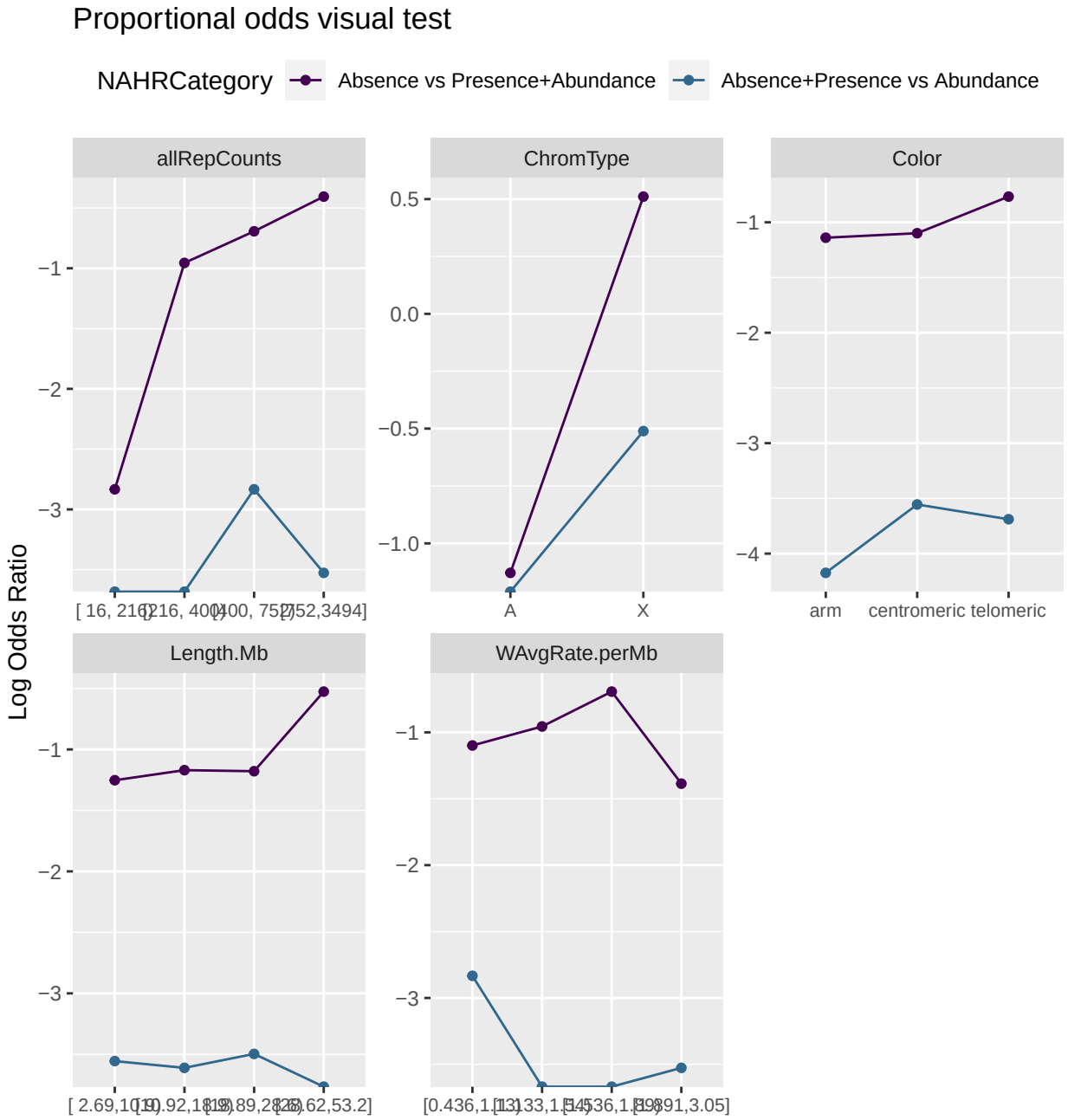
```
pacman::p_load("brant", "Hmisc")
btest<-brant(mod)
```

```
## ------------------------------------------------------
## Test for     X2  df  probability
## ------------------------------------------------------
```

```
## Omnibus          0   6   1
## Length.Mb        0   1   1
## allRepCounts     0   1   1
## Colorcentromeric 0   1   1
## Colortelomeric       0   1   1
## WAvgRate.perMb       0   1   1
## ChromTypeX       0   1   0.99
## ------------------------------------------------
##
## H0: Parallel Regression Assumption holds
```

	X2	df	probability
Omnibus	0.0001161	6	1.0000000
Length.Mb	0.0000013	1	0.9991042
allRepCounts	0.0000250	1	0.9960128
Colorcentromeric	0.0000000	1	0.9999230
Colortelomeric	0.0000000	1	0.9999387
WAvgRate.perMb	0.0000001	1	0.9997382
ChromTypeX	0.0001188	1	0.9913040

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.

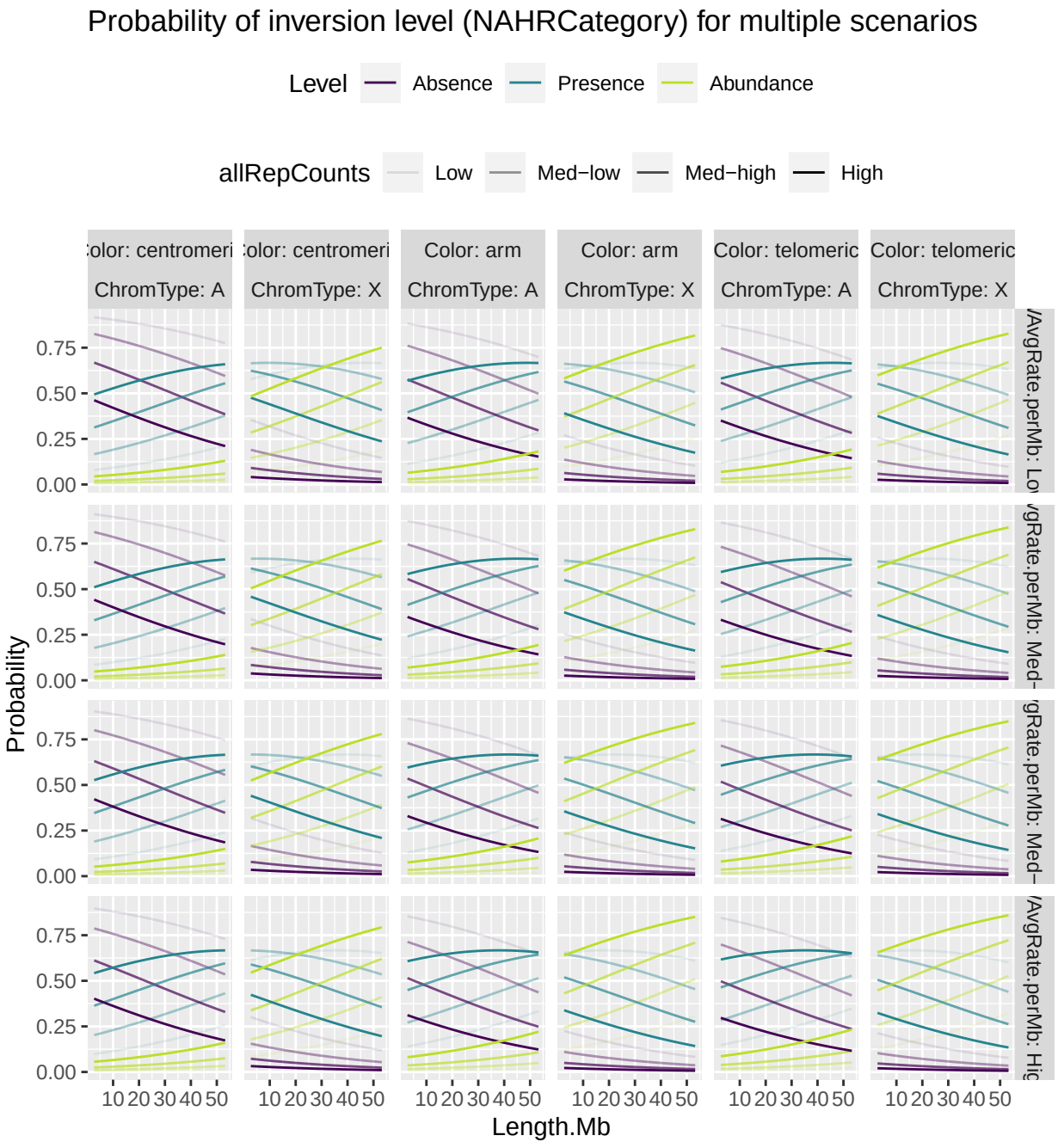


Figure 17: Probabiilty of having 0 to >3 inversions depending on multiple independent variables

Scaled variables

Total inversions model

Model fitting

```
## Call:
## polr(formula = myFormula, data = winRegions, Hess = T)
##
## Coefficients:
##                          Value Std. Error t value
## Length.Mb.Scaled        0.7380     0.2425  3.0434
## allRepCounts.Scaled     0.2002     0.1819  1.1008
## Colorcentromeric        0.3583     0.5872  0.6102
## Colortelomeric          0.1060     0.4751  0.2232
## WAvgRate.perMb.Scaled -0.1530     0.2664 -0.5745
## ChromTypeX              2.2375     0.8185  2.7337
##
## Intercepts:
##                     Value   Std. Error t value
## Absence|Presence   -0.0827  0.2646    -0.3125
## Presence|Abundance  2.7698  0.3866     7.1652
##
## Residual Deviance: 242.4534
## AIC: 258.4534
```

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	0.7379862	0.2424857	3.0434212	0.0023390
allRepCounts.Scaled	0.2002357	0.1819003	1.1007994	0.2709840
Colorcentromeric	0.3582987	0.5871806	0.6102018	0.5417282
Colortelomeric	0.1060310	0.4750690	0.2231907	0.8233871
WAvgRate.perMb.Scaled	-0.1530331	0.2663703	-0.5745124	0.5656211
ChromTypeX	2.2375283	0.8185035	2.7336822	0.0062630
Absence|Presence	-0.0826674	0.2645600	-0.3124711	0.7546825
Presence|Abundance	2.7698350	0.3865686	7.1651834	0.0000000

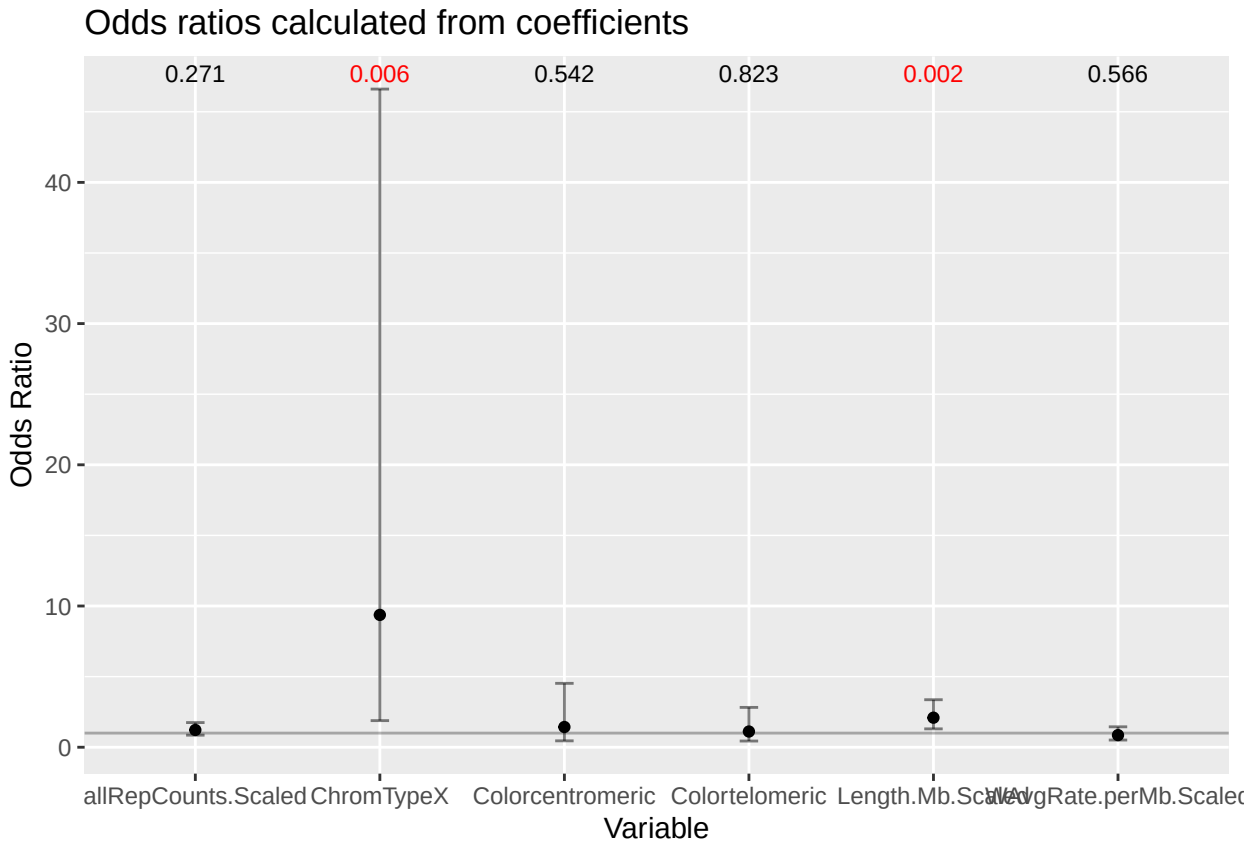
We can also get confidence intervals for the parameter estimates. These can be obtained either by profiling the likelihood function or by using the standard errors and assuming a normal distribution. Note that profiled CIs are not symmetric (although they are usually close to symmetric). If the 95% CI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

	2.5 %	97.5 %
Length.Mb.Scaled	0.2627229	1.2132495
allRepCounts.Scaled	-0.1562823	0.5567538
Colorcentromeric	-0.7925542	1.5091515
Colortelomeric	-0.8250872	1.0371492
WAvgRate.perMb.Scaled	-0.6751093	0.3690432
ChromTypeX	0.6332910	3.8417656

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	2.0917190	1.3004663	3.364400
allRepCounts.Scaled	1.2216907	0.8553177	1.744999
Colorcentromeric	1.4308929	0.4526871	4.522892
Colortelomeric	1.1118563	0.4381968	2.821163
WAvgRate.perMb.Scaled	0.8581014	0.5091008	1.446350
ChromTypeX	9.3701424	1.8838000	46.607691

Example of interpretation: "For 1 unit increase in Length.Mb.Scaled, a window is 2.091719 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.

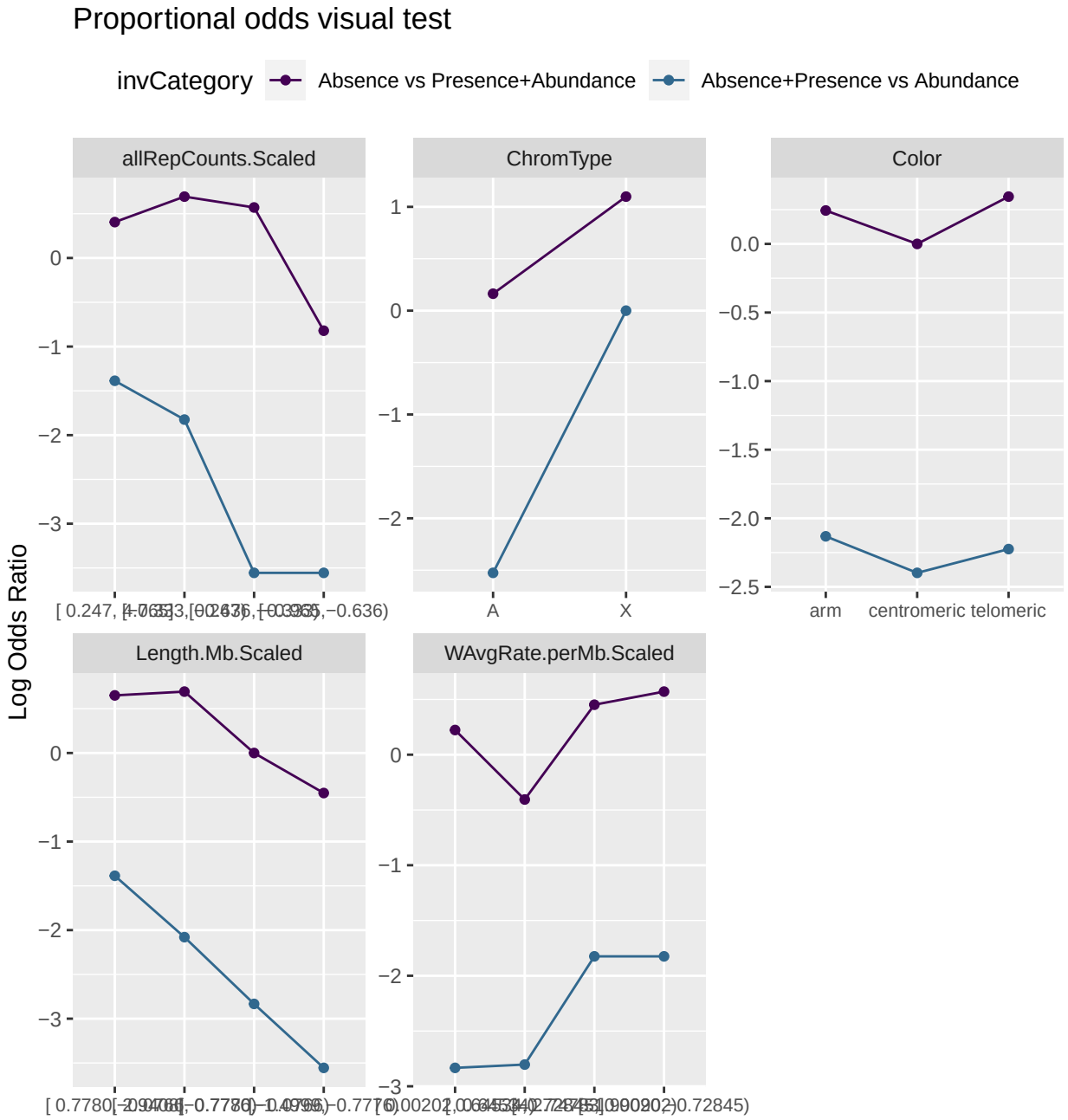
We test the parallel regression assumption with a Brant test:

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

```
## ------------------------------------------------------
## Test for     X2  df  probability
## ------------------------------------------------------
## Omnibus          10.79   6   0.09
## Length.Mb.Scaled 6.05    1   0.01
## allRepCounts.Scaled  0   1   0.96
## Colorcentromeric 2.1 1   0.15
## Colortelomeric       0.71    1   0.4
## WAvgRate.perMb.Scaled    0.42    1   0.52
## ChromTypeX       8.7 1   0
## ------------------------------------------------------
##
## H0: Parallel Regression Assumption holds
```

	X2	df	probability
Omnibus	10.7937767	6	0.0949630
Length.Mb.Scaled	6.0511470	1	0.0138973
allRepCounts.Scaled	0.0029780	1	0.9564800
Colorcentromeric	2.0995048	1	0.1473469
Colortelomeric	0.7113203	1	0.3990059
WAvgRate.perMb.Scaled	0.4197315	1	0.5170711
ChromTypeX	8.7025080	1	0.0031777

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.

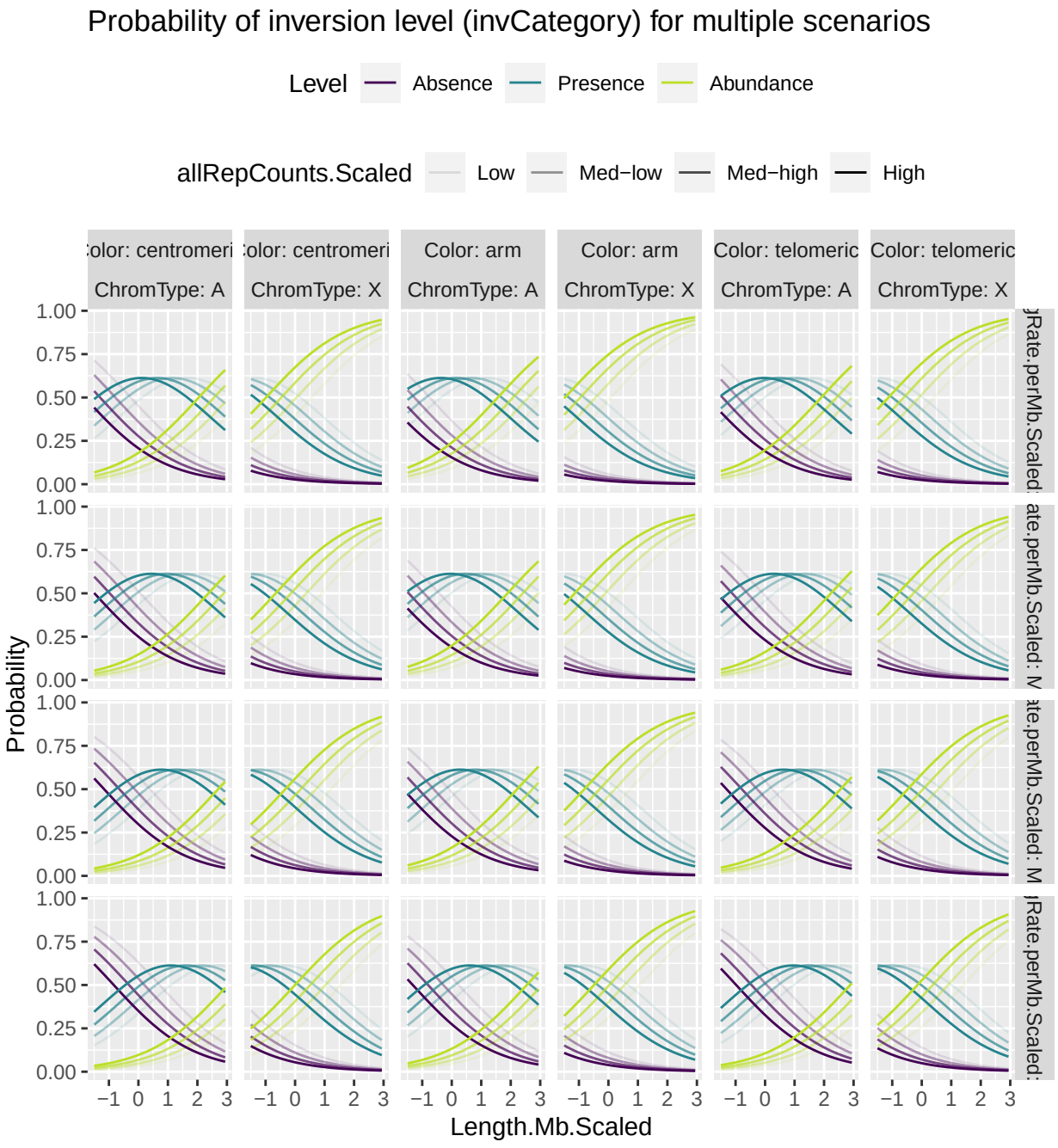


Figure 18: Probabiilty of having 0 to >3 inversions depending on multiple independent variables

NH inversions model

Model fitting

```
## Call:
## polr(formula = myFormula, data = winRegions, Hess = T)
##
## Coefficients:
##                          Value Std. Error  t value
## Length.Mb.Scaled       1.00332     0.2703  3.71198
## allRepCounts.Scaled   -0.26993     0.2026 -1.33234
## Colorcentromeric       0.18498     0.6326  0.29241
## Colortelomeric         0.02217     0.5424  0.04089
## WAvgRate.perMb.Scaled -0.30916     0.3204 -0.96483
## ChromTypeX            -0.85310     0.8894 -0.95915
##
## Intercepts:
##                     Value   Std. Error t value
## Absence|Presence    0.5251  0.2889     1.8175
## Presence|Abundance  3.8049  0.5461     6.9674
##
## Residual Deviance: 195.6799
## AIC: 211.6799
```

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	1.0033218	0.2702929	3.7119791	0.0002056
allRepCounts.Scaled	-0.2699262	0.2025949	-1.3323444	0.1827471
Colorcentromeric	0.1849802	0.6326003	0.2924125	0.7699713
Colortelomeric	0.0221747	0.5423666	0.0408851	0.9673875
WAvgRate.perMb.Scaled	-0.3091573	0.3204275	-0.9648274	0.3346313
ChromTypeX	-0.8531026	0.8894387	-0.9591472	0.3374846
Absence|Presence	0.5250946	0.2889145	1.8174736	0.0691446
Presence|Abundance	3.8048703	0.5460970	6.9673895	0.0000000

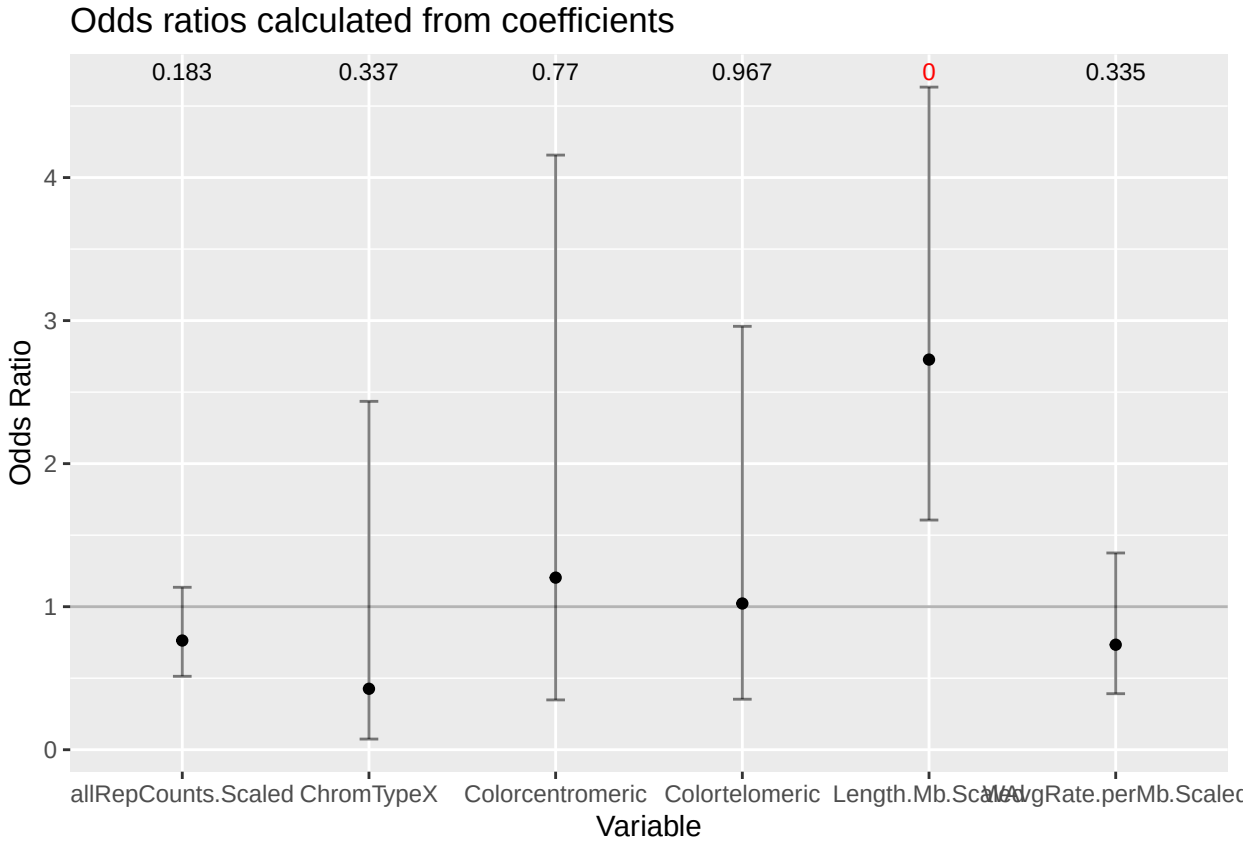
We can also get confidence intervals for the parameter estimates. These can be obtained either by profiling the likelihood function or by using the standard errors and assuming a normal distribution. Note that profiled CIs are not symmetric (although they are usually close to symmetric). If the 95% CI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

	2.5 %	97.5 %
Length.Mb.Scaled	0.4735573	1.5330862
allRepCounts.Scaled	-0.6670049	0.1271525
Colorcentromeric	-1.0548936	1.4248541
Colortelomeric	-1.0408443	1.0851937
WAvgRate.perMb.Scaled	-0.9371837	0.3188692
ChromTypeX	-2.5963704	0.8901651

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	2.7273264	1.6056960	4.632452
allRepCounts.Scaled	0.7634358	0.5132435	1.135590
Colorcentromeric	1.2031947	0.3482295	4.157251
Colortelomeric	1.0224224	0.3531564	2.960013
WAvgRate.perMb.Scaled	0.7340653	0.3917295	1.375571
ChromTypeX	0.4260909	0.0745437	2.435532

Example of interpretation: "For 1 unit increase in Length.Mb.Scaled, a window is 2.7273264 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:

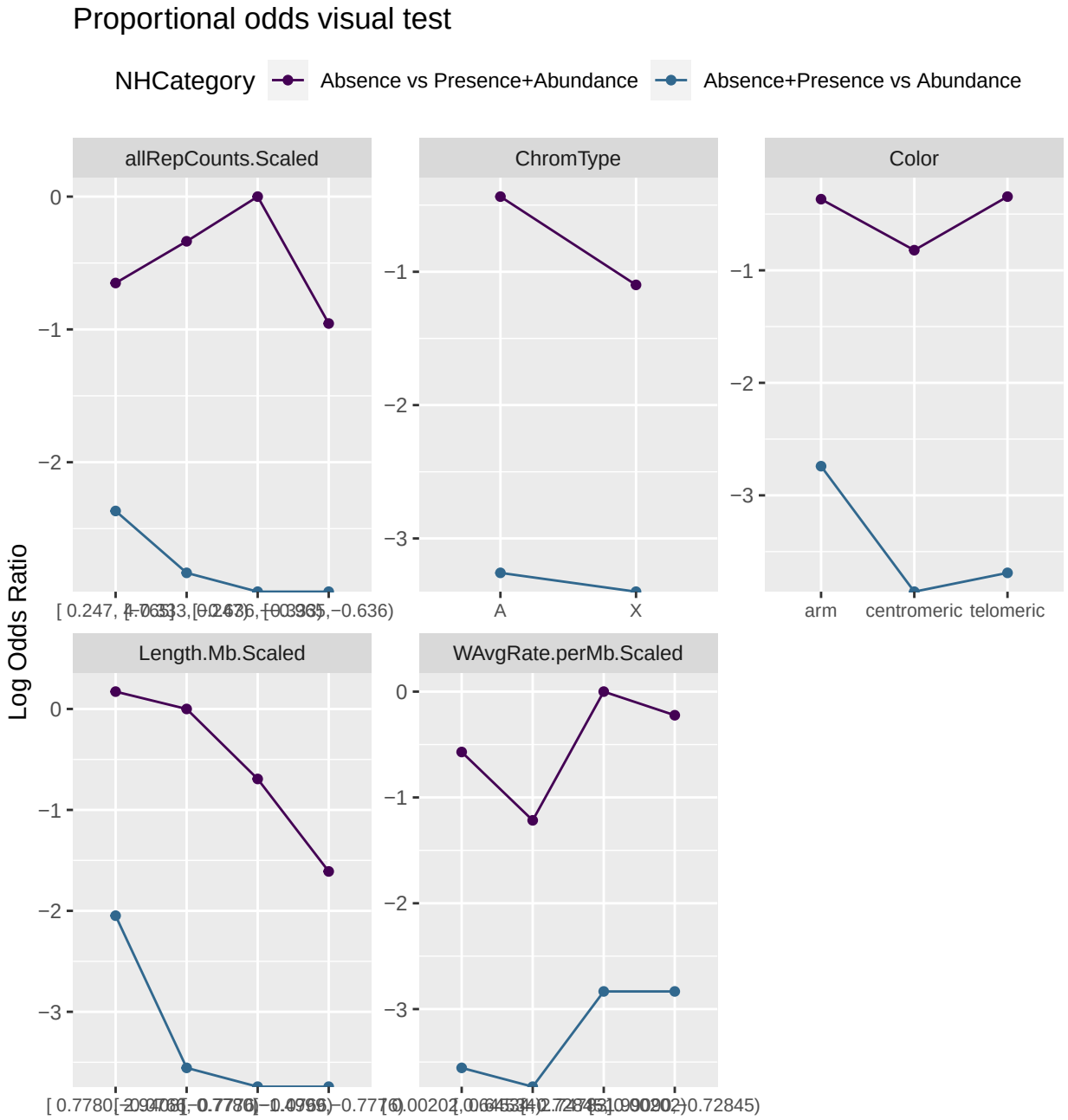
```
pacman::p_load("brant", "Hmisc")
btest<-brant(mod)
```

```
## ------------------------------------------------------
## Test for     X2  df  probability
## ------------------------------------------------------
```

```
## Omnibus          4.34    6   0.63
## Length.Mb.Scaled 2.34    1   0.13
## allRepCounts.Scaled  0.37    1   0.55
## Colorcentromeric 0   1   1
## Colortelomeric       0.78    1   0.38
## WAvgRate.perMb.Scaled    0.01    1   0.93
## ChromTypeX       0   1   1
## ----------------------------------------------------
##
## H0: Parallel Regression Assumption holds
```

	X2	df	probability
Omnibus	4.3357879	6	0.6313347
Length.Mb.Scaled	2.3371907	1	0.1263172
allRepCounts.Scaled	0.3662969	1	0.5450297
Colorcentromeric	0.0000132	1	0.9971010
Colortelomeric	0.7810778	1	0.3768117
WAvgRate.perMb.Scaled	0.0079667	1	0.9288782
ChromTypeX	0.0000033	1	0.9985428

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.

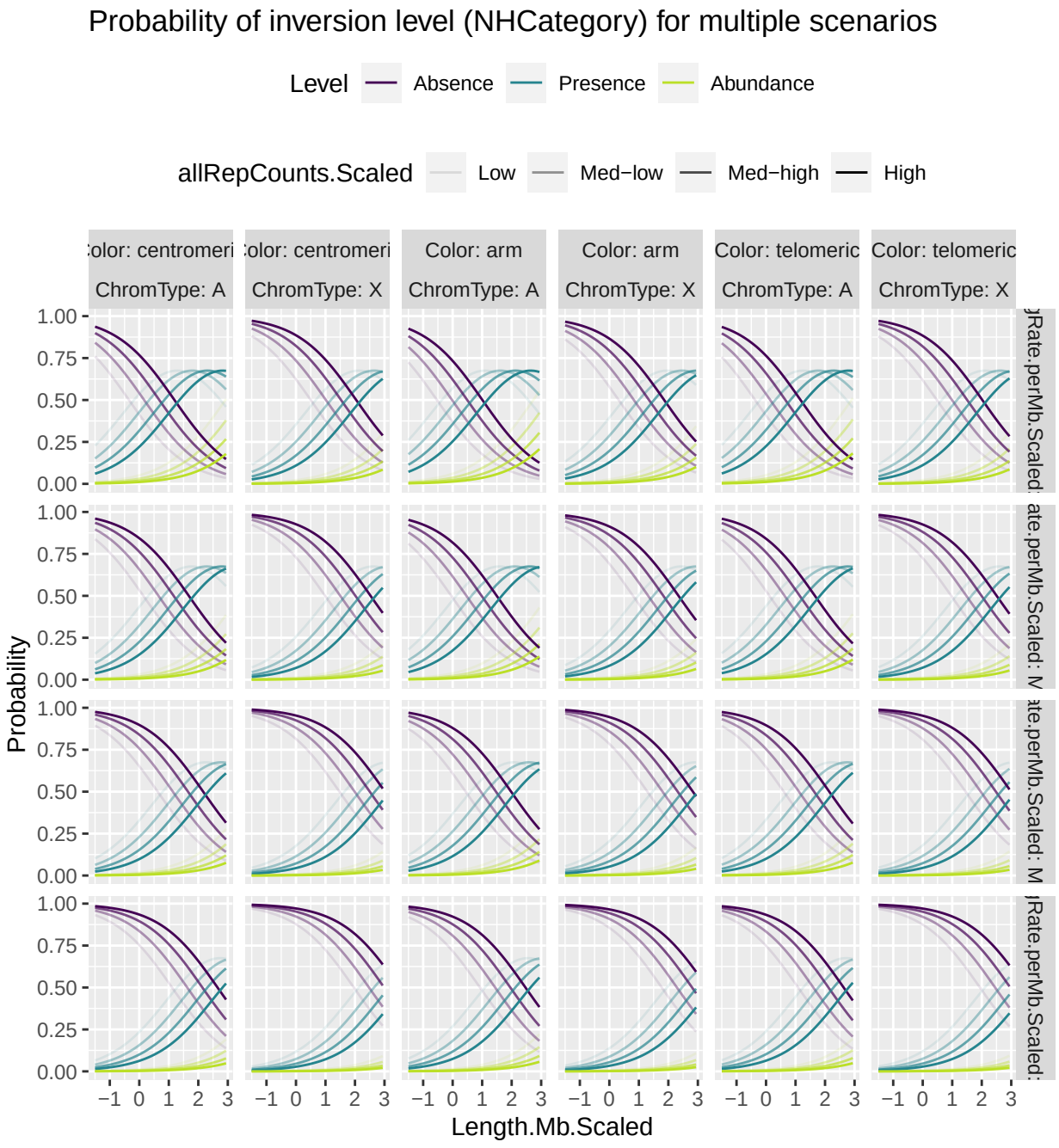


Figure 19: Probabiilty of having 0 to >3 inversions depending on multiple independent variables

NAHR inversions model

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

Model fitting

```
## Call:
## polr(formula = myFormula, data = winRegions, Hess = T)
##
## Coefficients:
##                      Value Std. Error t value
## Length.Mb        0.0231024   0.024028  0.9615
## allRepCounts     0.0007353   0.000369  1.9928
## Colorcentromeric 0.3953894   0.726258  0.5444
## Colortelomeric   0.4640769   0.555281  0.8358
## WAvgRate.perMb   0.0934910   0.592376  0.1578
## ChromTypeX       3.0023034   0.916542  3.2757
##
## Intercepts:
##                     Value  Std. Error t value
## Absence|Presence   2.5192 1.3241     1.9026
## Presence|Abundance 5.7408 1.4843     3.8676
##
## Residual Deviance: 166.7064
## AIC: 182.7064
```

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

	Value	Std. Error	t value	p value
Length.Mb	0.0231024	0.0240280	0.9614763	0.3363128
allRepCounts	0.0007353	0.0003690	1.9927722	0.0462864
Colorcentromeric	0.3953894	0.7262578	0.5444203	0.5861523
Colortelomeric	0.4640769	0.5552813	0.8357511	0.4032949
WAvgRate.perMb	0.0934910	0.5923755	0.1578239	0.8745956
ChromTypeX	3.0023034	0.9165423	3.2756844	0.0010541
Absence|Presence	2.5192498	1.3241418	1.9025529	0.0570989
Presence|Abundance	5.7408086	1.4843305	3.8676081	0.0001099

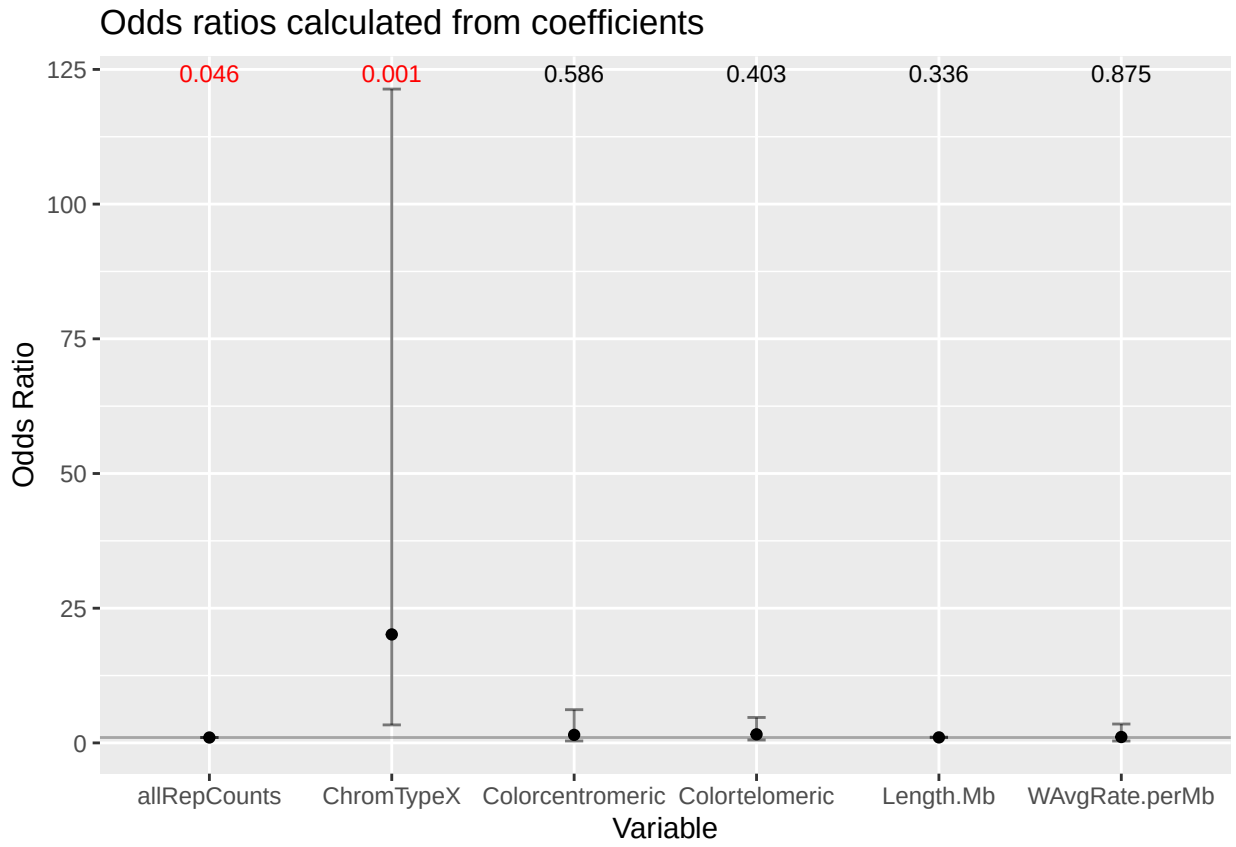
We can also get confidence intervals for the parameter estimates. These can be obtained either by profiling the likelihood function or by using the standard errors and assuming a normal distribution. Note that profiled CIs are not symmetric (although they are usually close to symmetric). If the 95% CI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

	2.5 %	97.5 %
Length.Mb	-0.0239917	0.0701964
allRepCounts	0.0000121	0.0014586
Colorcentromeric	-1.0280496	1.8188285
Colortelomeric	-0.6242543	1.5524082
WAvgRate.perMb	-1.0675437	1.2545257
ChromTypeX	1.2059135	4.7986934

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	1.023371	0.9762938	1.072719
allRepCounts	1.000736	1.0000121	1.001460
Colorcentromeric	1.484962	0.3577039	6.164632
Colortelomeric	1.590545	0.5356607	4.722830
WAvgRate.perMb	1.098001	0.3438521	3.506175
ChromTypeX	20.131856	3.3398085	121.351755

Example of interpretation: "For 1 unit increase in Length.Mb, a window is 1.0233713 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.

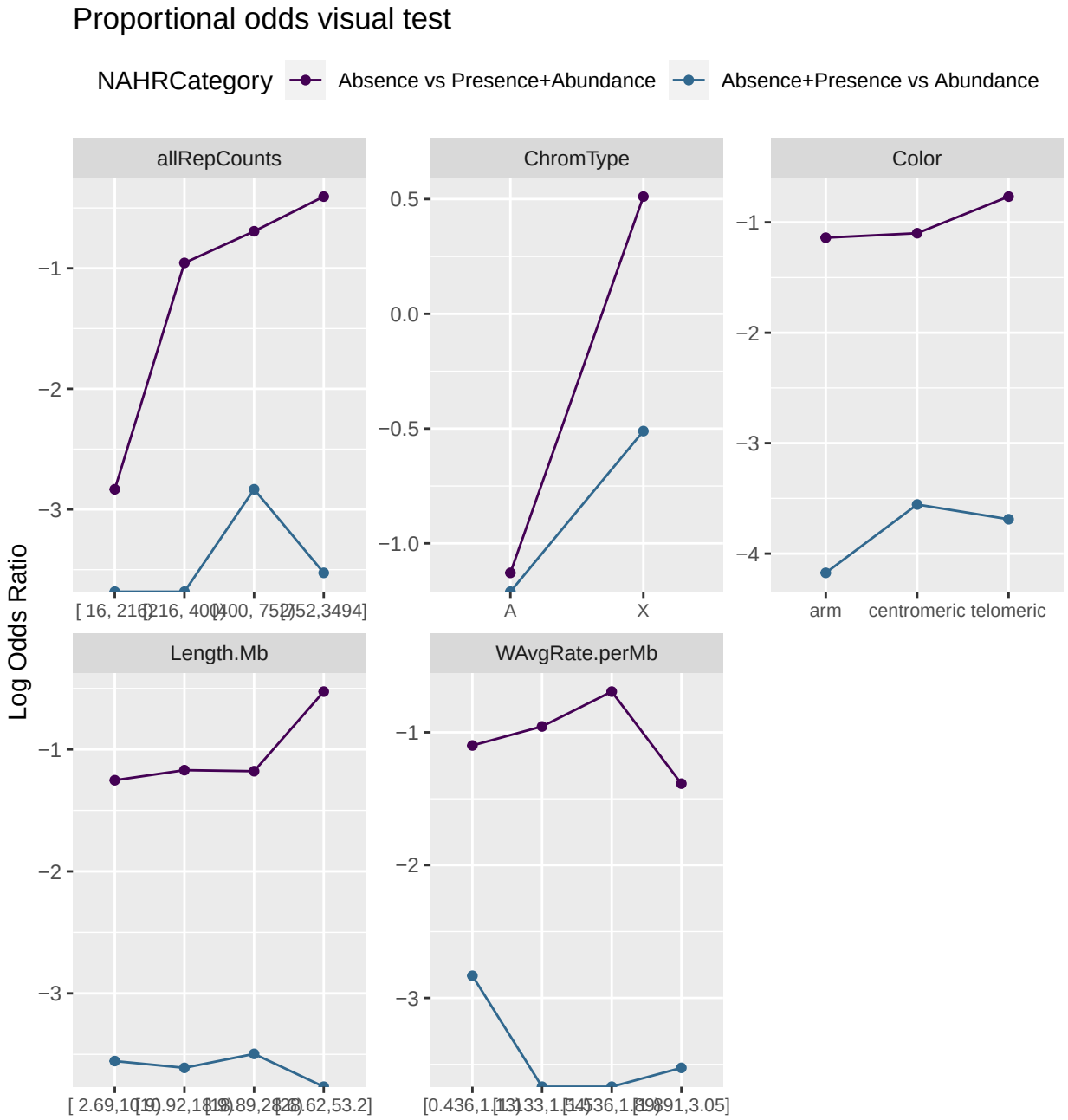
We test the parallel regression assumption with a Brant test:

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

```
## ------------------------------------------------------
## Test for     X2  df  probability
## ------------------------------------------------------
## Omnibus          0   6   1
## Length.Mb        0   1   1
## allRepCounts     0   1   1
## Colorcentromeric 0   1   1
## Colortelomeric       0   1   1
## WAvgRate.perMb       0   1   1
## ChromTypeX       0   1   0.99
## ------------------------------------------------------
##
## H0: Parallel Regression Assumption holds
```

	X2	df	probability
Omnibus	0.0001161	6	1.0000000
Length.Mb	0.0000013	1	0.9991042
allRepCounts	0.0000250	1	0.9960128
Colorcentromeric	0.0000000	1	0.9999230
Colortelomeric	0.0000000	1	0.9999387
WAvgRate.perMb	0.0000001	1	0.9997382
ChromTypeX	0.0001188	1	0.9913040

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.

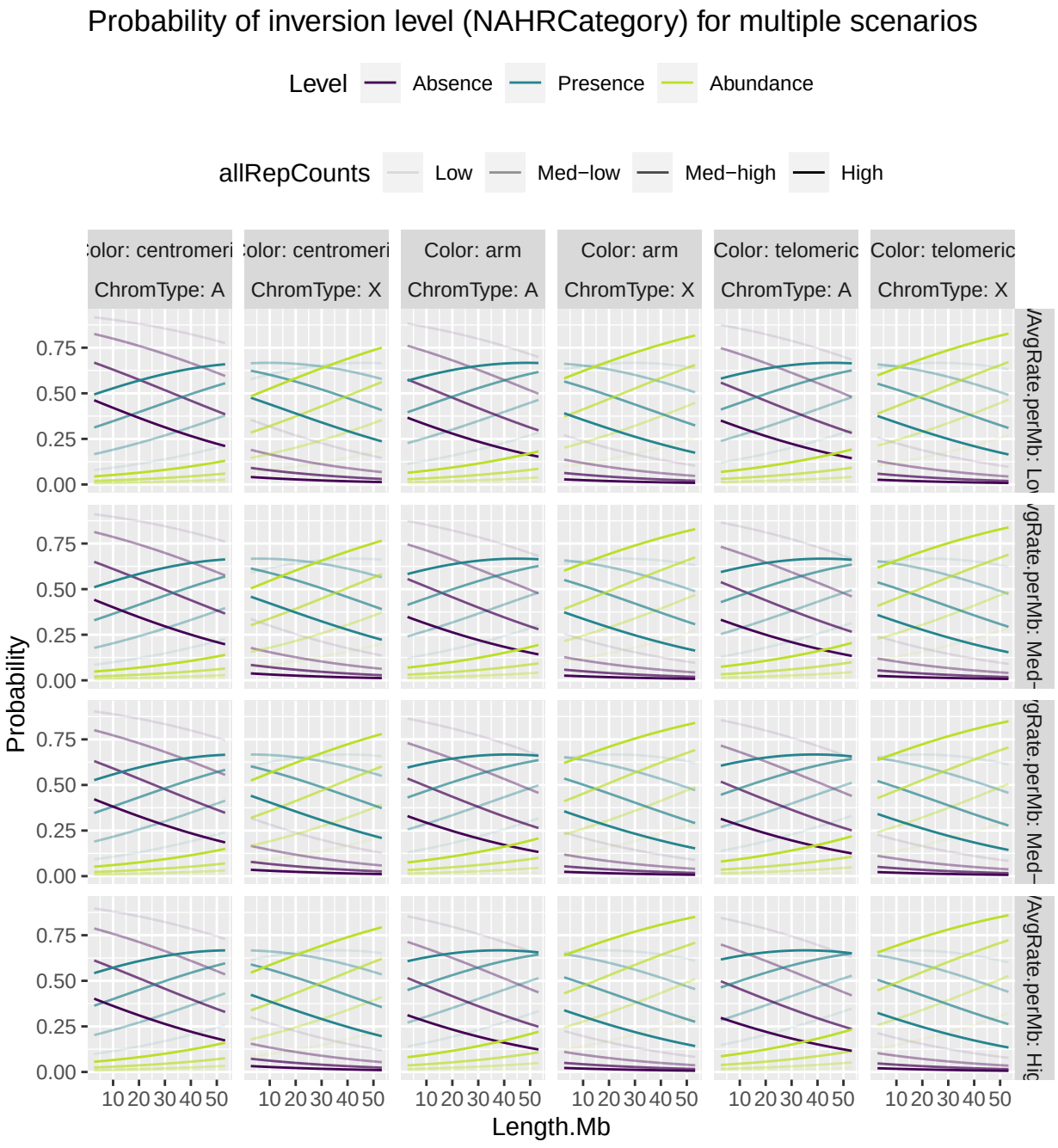


Figure 20: Probabiilty of having 0 to >3 inversions depending on multiple independent variables