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.

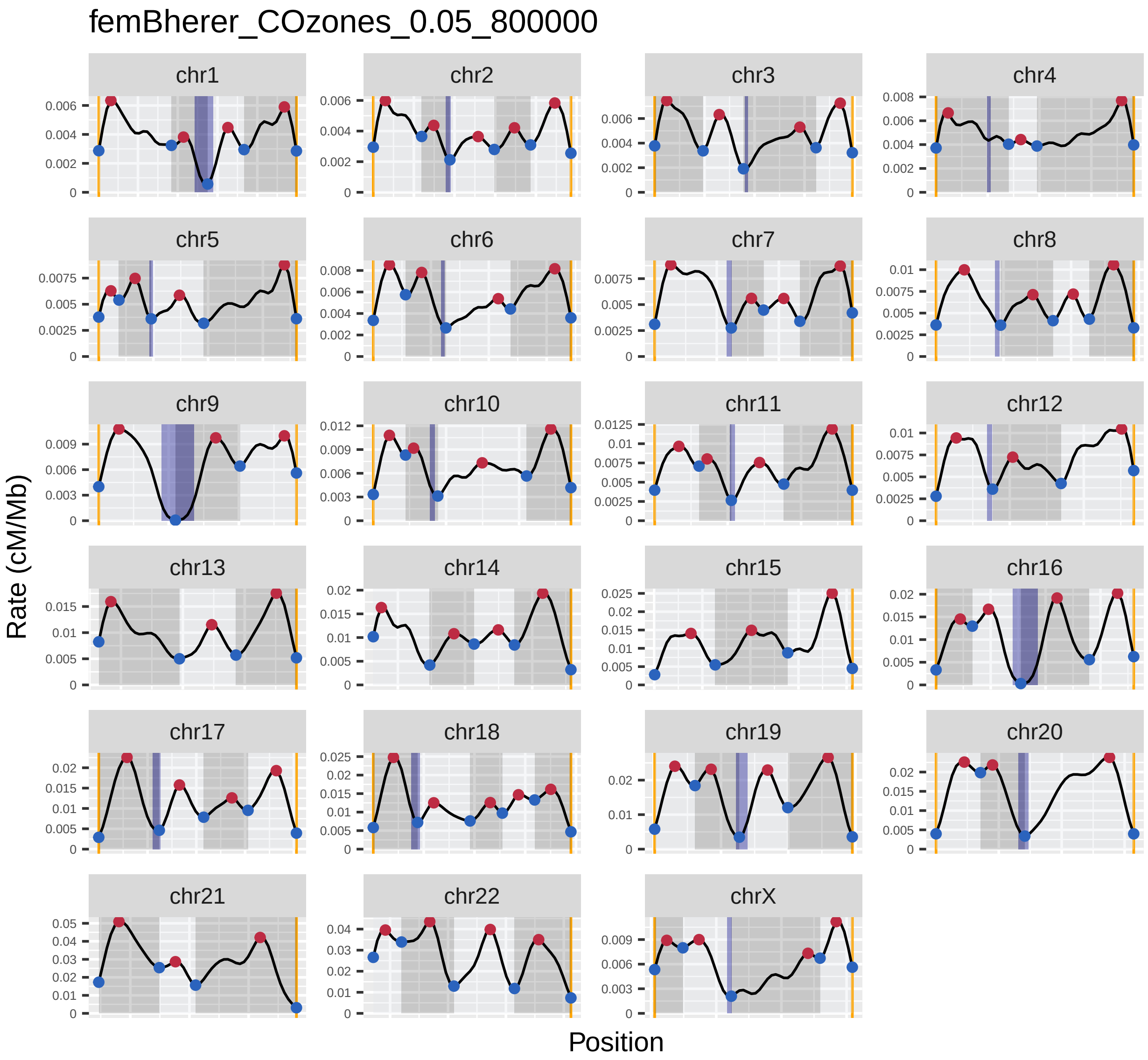


Figure 1: Crossover zones; centromeres in blue, workspace limits in orange.

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.

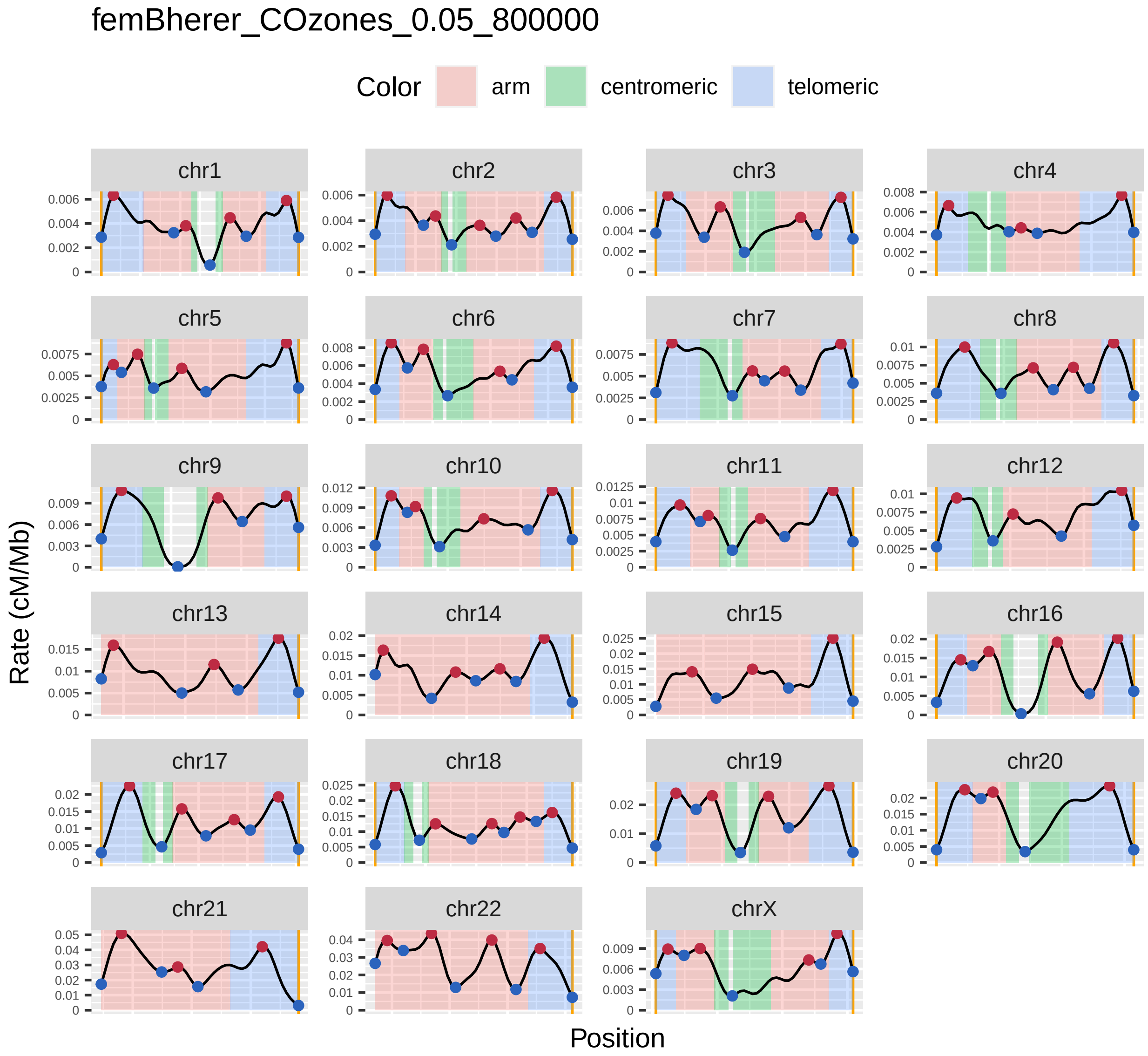


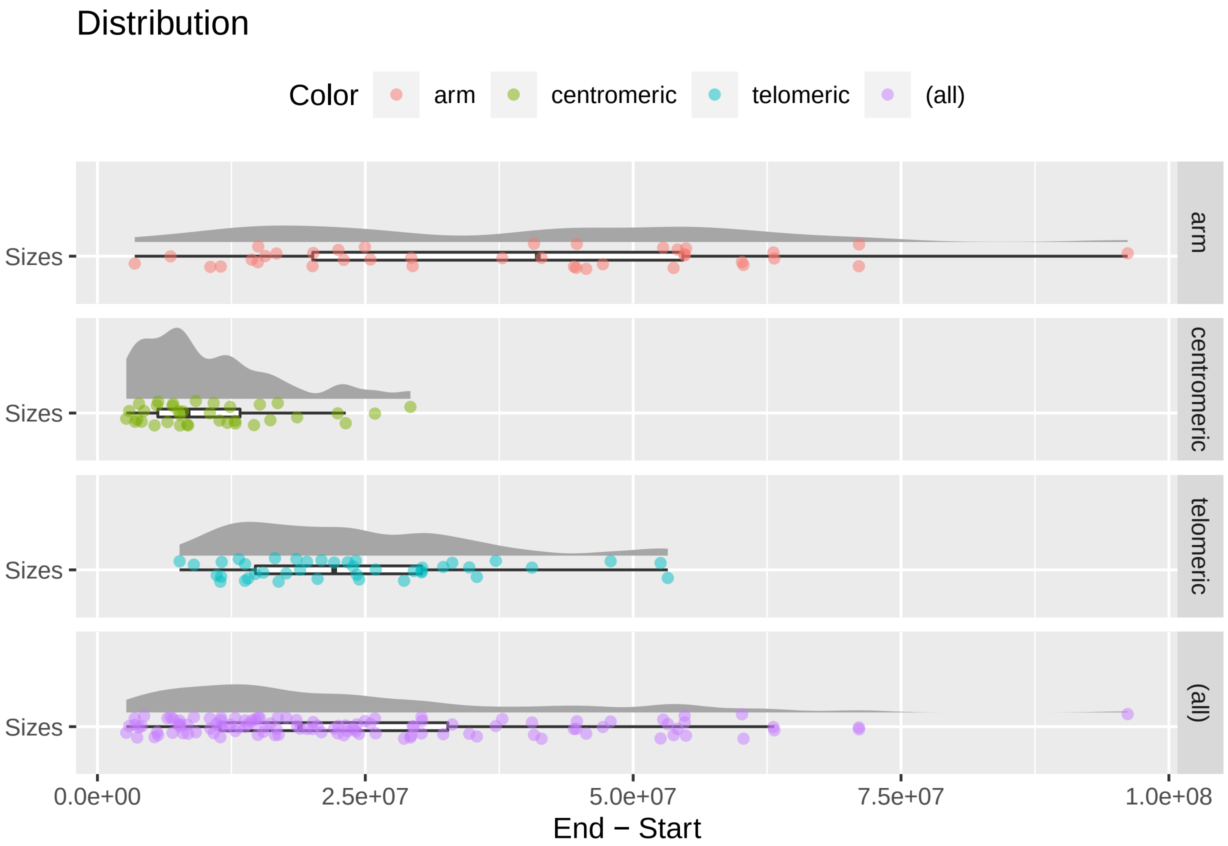
Figure 2: Color-coded windows for telomeric, centromeric and arm categories.

```
pacman::p_load(ggdist, ggplot2, gghalves, reshape2, patchwork)

#Look for size outliers
sizes<-windows$End-windows$Start

ggplot(windows, aes(y = "Sizes",  End-Start))+
      # Half violin
```

```
    ggdist::stat_halfeye(adjust = .5, width = .6, .width = 0, justification = -.2, point_colour = NA)
    # Boxplot
    geom_boxplot(width = .1, outlier.shape = NA) +
    # Points
    gghalves::geom_half_point_panel(side = "l", range_scale = .6,  alpha = .5, aes(color = Color))+
    # scale_color_manual(values = c(rep("#3c7ae7",11),rep("#89b23e",11) ))+
    # Adjust coordinates
    # coord_flip()+
    # coord_flip( xlim = c(1.3, NA))+
    # Adjust labels
    theme(axis.title.y = element_blank(), legend.position = "top")+
    facet_grid(Color~., margins=TRUE)+
    # Title
    ggtitle("Distribution")
```



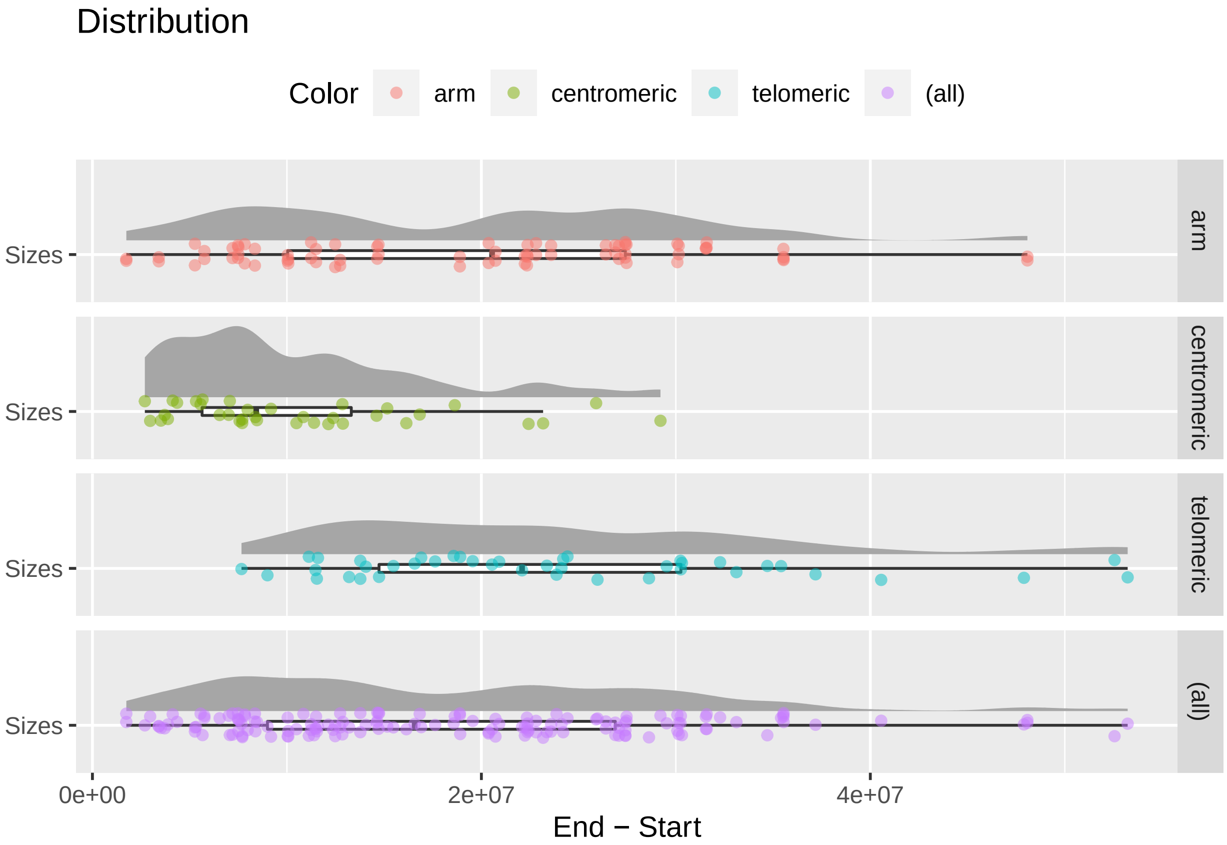
```
# So what happend if I divide each "arm" window in 2
windows_telocen<-windows[windows$Color != "arm",]
windows_arm<-windows[windows$Color == "arm",]

windows_arm$half<-(windows_arm$End+windows_arm$Start)/2
windows_arm_alpha<-windows_arm[,c("Start", "half", "Chromosome", "Color")]
windows_arm_beta<-windows_arm[,c("half", "End", "Chromosome", "Color")]

colnames(windows_arm_alpha)<-colnames(windows)
colnames(windows_arm_beta)<-colnames(windows)

windows_halved<-rbind(rbind(windows_telocen, windows_arm_alpha), windows_arm_beta)
```

```r
ggplot(windows_halved, aes(y = "Sizes",  End-Start))+
      # Half violin
      ggdist::stat_halfeye(adjust = .5, width = .6, .width = 0, justification = -.2, point_colour = NA)
      # Boxplot
      geom_boxplot(width = .1, outlier.shape = NA) +
      # Points
      gghalves::geom_half_point_panel(side = "l", range_scale = .6,  alpha = .5, aes(color = Color))+
      # scale_color_manual(values = c(rep("#3c7ae7",11),rep("#89b23e",11) ))+
      # Adjust coordinates
      # coord_flip()+
      # coord_flip( xlim = c(1.3, NA))+
      # Adjust labels
      theme(axis.title.y = element_blank(), legend.position = "top")+
      facet_grid(Color~., margins=TRUE)+
      # Title
      ggtitle("Distribution")
```



```r
write.table(windows_halved, "data/windows.txt", quote = F, row.names = F, col.names = T, sep = "\t")
```

–> –> –>

–> –>

–> –> –>

–> –>

–> –> –>

–> –>

–> –> –>

–> –>

–> –> –>

–> –>

–> –> –>

–> –>

\title(With the X)

Numerical categories

Descriptive statistics

Raw data:

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

For each window, I calculated the number of total inversions, NH inversions, and NAHR inversions, the window length in Mb, number of repeats and the average recombination rate in cM/Mb.

I want to perform Ordinal Logistic Regressions on different subsets of the data. The assumptions of the Ordinal Logistic Regression are as follow:

1. The dependent variable is ordered.
2. One or more of the independent variables are either continuous, categorical or ordinal.
3. No multi-collinearity.
4. Proportional odds.

I show the data distributions in the figure below. The inversion counts have only a number of possible options, so they can be considered an ordinal variable. The independent variables are continuous and categorical, so assumptions 1 and 2 are satisfied

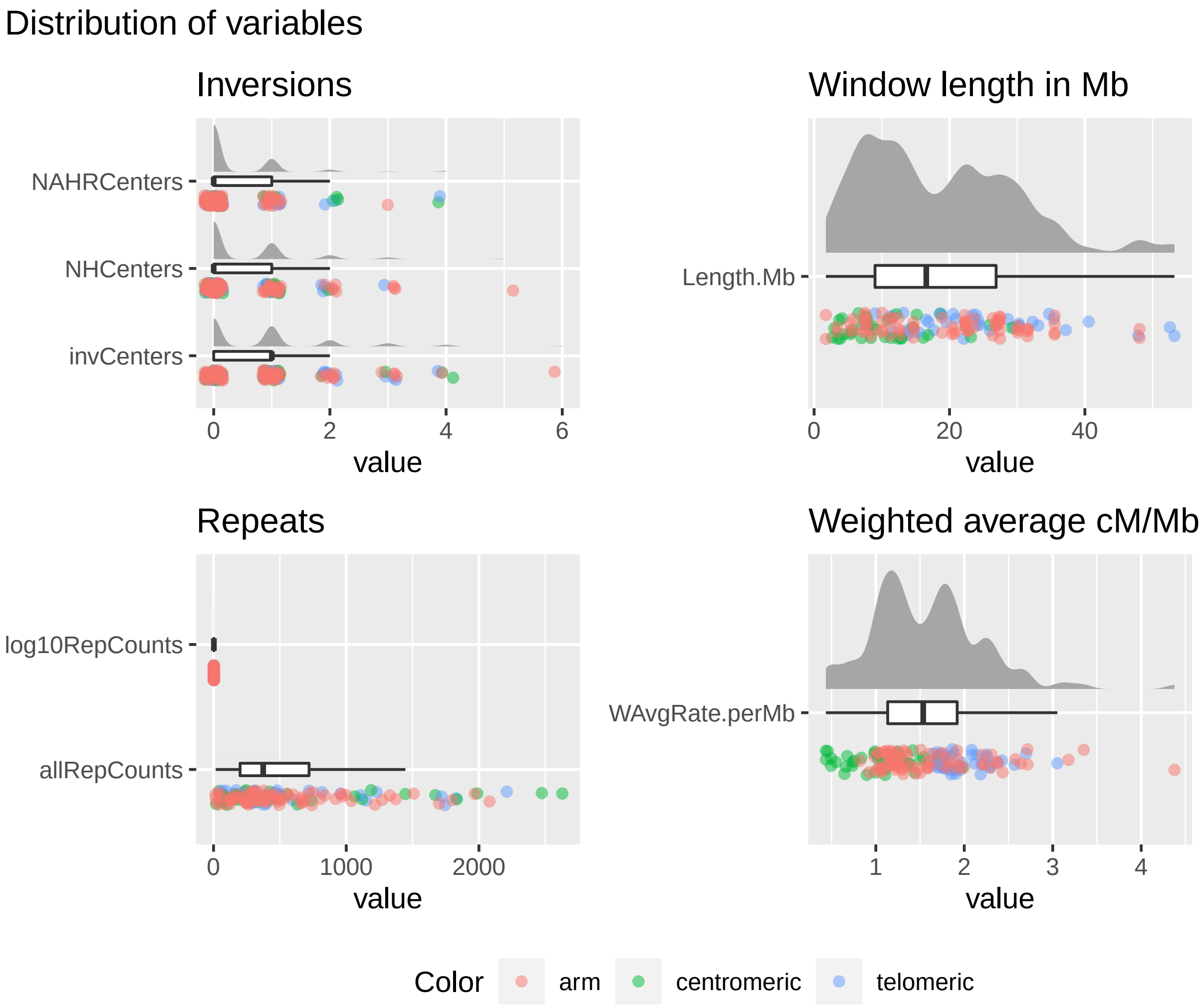


Figure 3: Distribution of variables.

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

Table 2: Original counts

CountGroups	invCenters	NHCenters	NAHRCenters
0	72	97	114
1	52	41	31
2	16	10	5
3	8	4	1
4	4	NA	2
5	NA	1	NA
6	1	NA	NA

Table 3: New counts

CountGroups	invCategory	NHCategory	NAHRCategory
0	72	97	114

CountGroups	invCategory	NHCategory	NAHRCategory
1	52	41	31
2	16	10	5
3+	13	5	3

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

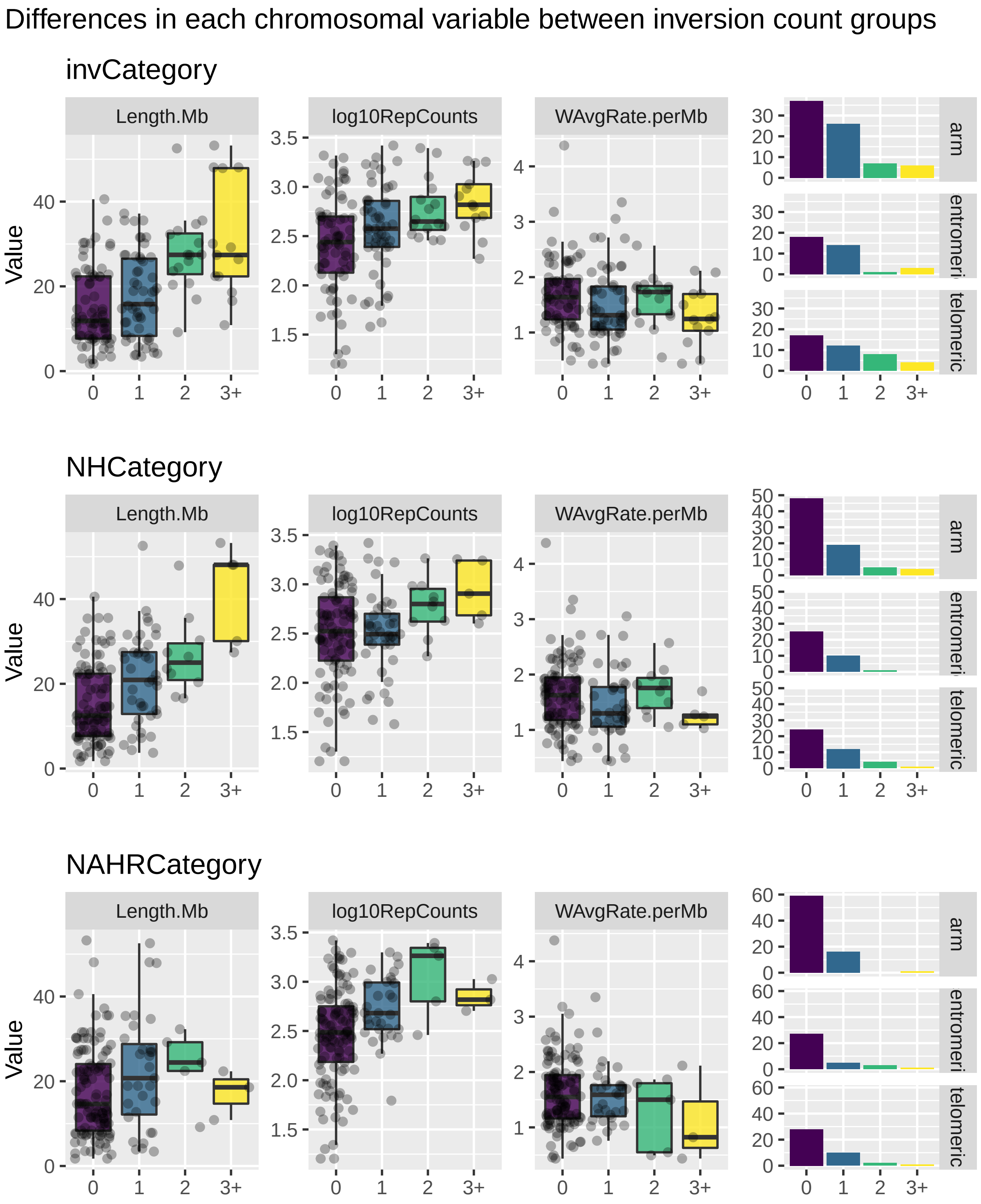


Figure 4: Potential effect of independent variables on the different types of invesions.

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

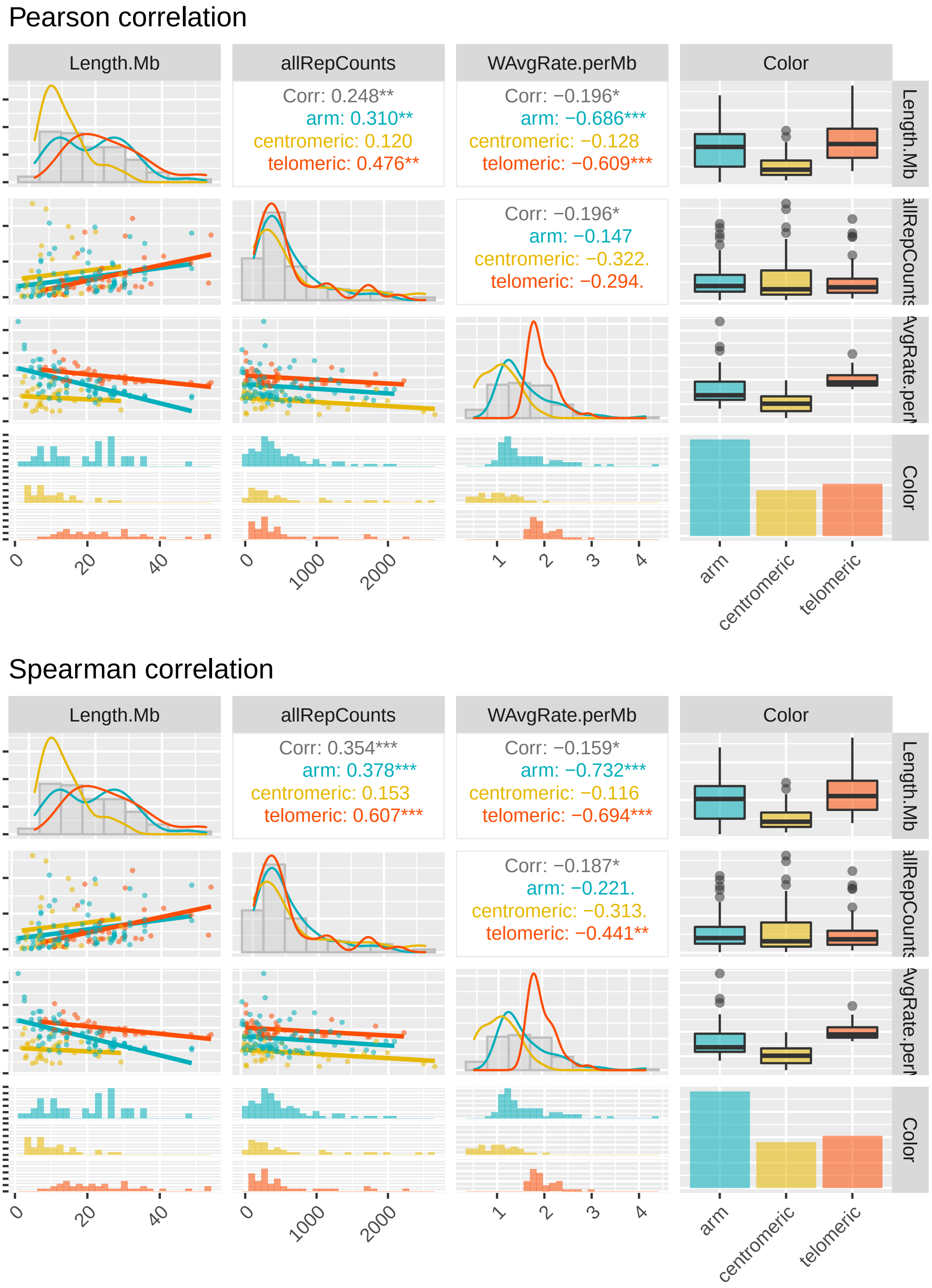


Figure 5: Correlations between variables.

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	1.931714	1	1.389861
allRepCounts	1.105951	1	1.051642
Color	2.573944	2	1.266630
WAvgRate.perMb	2.163202	1	1.470783

	GVIF	Df	GVIF^(1/(2*Df))
scale(Length.Mb)	1.931714	1	1.389861
scale(allRepCounts)	1.105951	1	1.051642
Color	2.573944	2	1.266630
scale(WAvgRate.perMb)	2.163202	1	1.470783

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.	1.741944	-1.4979278	16.0000	-0.9932976	0.4356883	-1.8539902
1st Qu.	8.999548	-0.8481517	200.0000	-0.6589042	1.1341848	-0.7230993
Median	16.569122	-0.1704450	374.0000	-0.3426843	1.5359258	-0.0726664
Mean	18.472889	0.0000000	562.5621	0.0000000	1.5808082	0.0000000
3rd Qu.	26.886669	0.7532889	720.0000	0.2861206	1.9196997	0.5486773
Max.	53.232426	3.1120343	2628.0000	3.7536347	4.3762818	4.5259719

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       0.0914777  0.0208880 4.37945
## allRepCounts    0.0005871  0.0003134 1.87294
## Colorcentromeric 0.7302445  0.5431841 1.34438
## Colortelomeric  0.0319482  0.4483824 0.07125
## WAvgRate.perMb  0.1165490  0.4155894 0.28044
## ChromTypeX      2.2468350  0.7882983 2.85023
##
## Intercepts:
##      Value  Std. Error t value
## 0|1  2.2407 0.9918     2.2592
## 1|2  4.1867 1.0349     4.0455
## 2|3+ 5.3926 1.0987     4.9083
##
## Residual Deviance: 311.8409
## AIC: 329.8409
```

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

	Value	Std. Error	t value	p value
Length.Mb	0.0914777	0.0208880	4.3794481	0.0000119
allRepCounts	0.0005871	0.0003134	1.8729417	0.0610764
Colorcentromeric	0.7302445	0.5431841	1.3443775	0.1788264
Colortelomeric	0.0319482	0.4483824	0.0712521	0.9431971
WAvgRate.perMb	0.1165490	0.4155894	0.2804426	0.7791380
ChromTypeX	2.2468350	0.7882983	2.8502346	0.0043687
0|1	2.2406887	0.9918176	2.2591741	0.0238726
1|2	4.1866799	1.0348960	4.0455078	0.0000522
2|3+	5.3926193	1.0986696	4.9083176	0.0000009

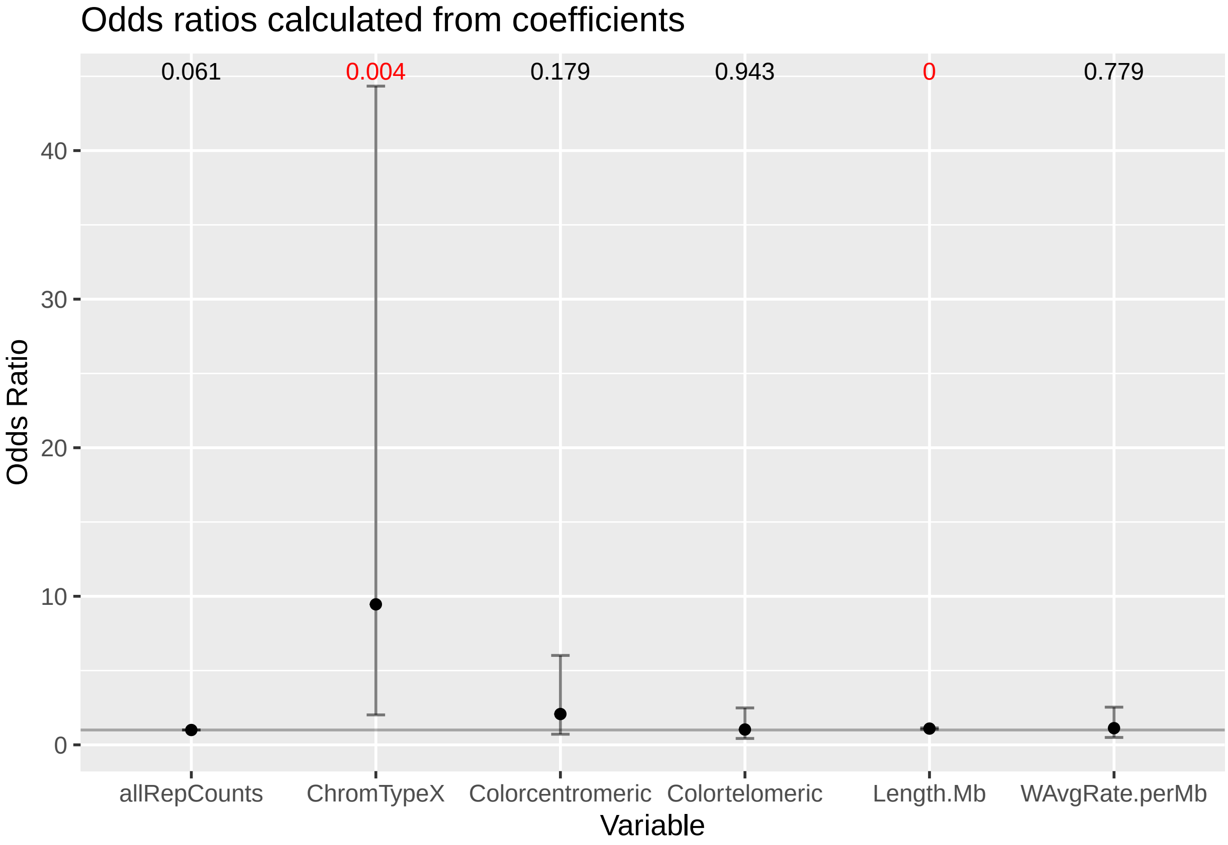
We can also get confidence intervals for the parameter estimates. These can be obtained either by profiling the likelihood function or by using the standard errors and assuming a normal distribution. Note that profiled CIs are not symmetric (although they are usually close to symmetric). If the 95% CI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

	2.5 %	97.5 %
Length.Mb	0.0505381	0.1324173
allRepCounts	-0.0000273	0.0012014
Colorcentromeric	-0.3343768	1.7948659
Colortelomeric	-0.8468651	0.9107615
WAvgRate.perMb	-0.6979914	0.9310893
ChromTypeX	0.7017988	3.7918713

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	1.095792	1.0518369	1.141585
allRepCounts	1.000587	0.9999727	1.001202
Colorcentromeric	2.075588	0.7157840	6.018668
Colortelomeric	1.032464	0.4287569	2.486215
WAvgRate.perMb	1.123612	0.4975838	2.537272
ChromTypeX	9.457755	2.0173782	44.339293

Example of interpretation: "For 1 unit increase in Length.Mb, a window is 1.0957923 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.

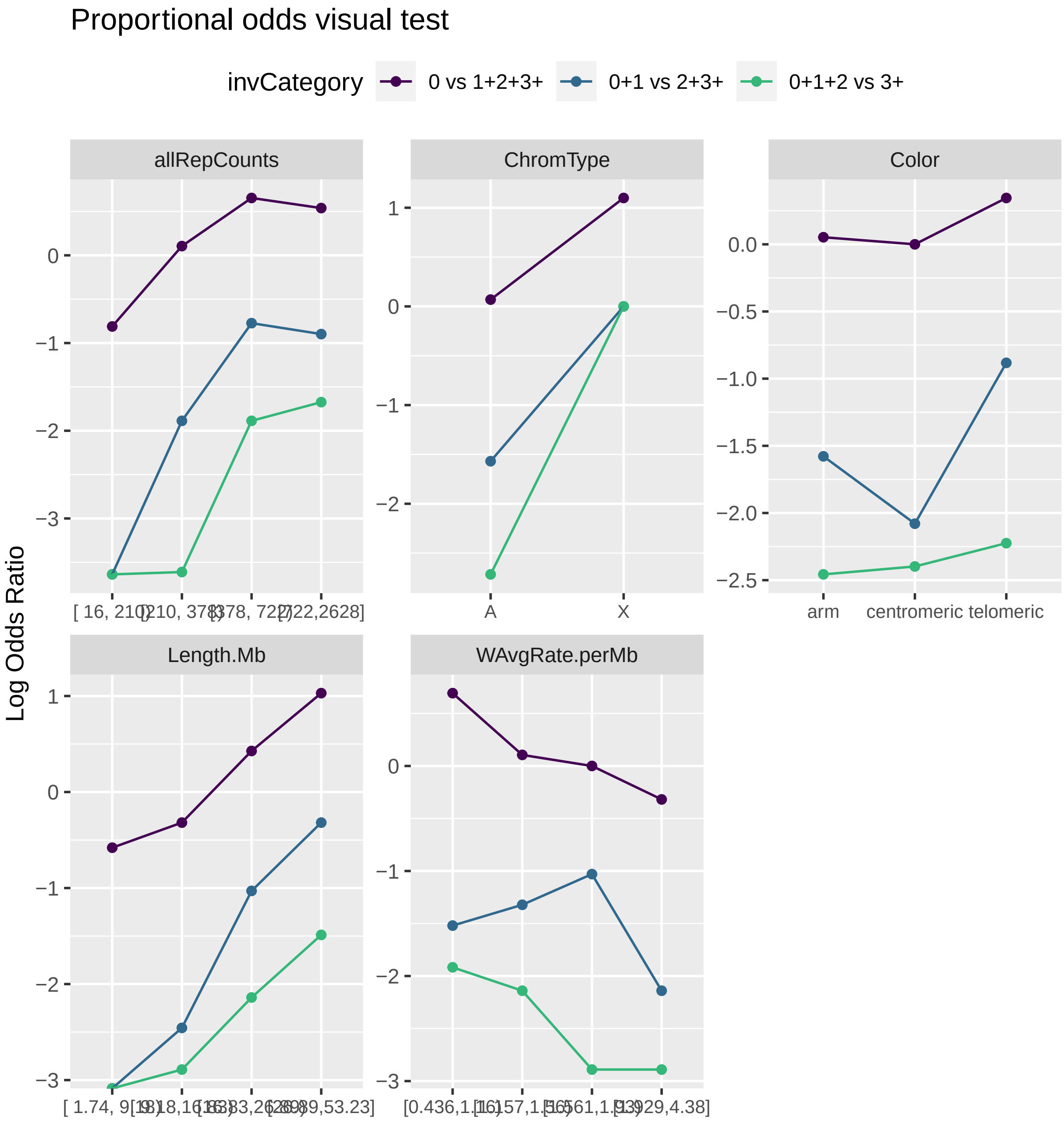
We test the parallel regression assumption with a Brant test:

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

```
## --------------------------------------------------------
## Test for     X2  df  probability
## --------------------------------------------------------
## Omnibus          42.36   12  0
## Length.Mb        5.32    2   0.07
## allRepCounts     0.1 2   0.95
## Colorcentromeric 1.71    2   0.43
## Colortelomeric       1.05    2   0.59
## WAvgRate.perMb       0.47    2   0.79
## ChromTypeX       10.04   2   0.01
## --------------------------------------------------------
##
## H0: Parallel Regression Assumption holds
```

	X2	df	probability
Omnibus	42.3581930	12	0.0000290
Length.Mb	5.3215725	2	0.0698932
allRepCounts	0.0978698	2	0.9522431
Colorcentromeric	1.7067240	2	0.4259804
Colortelomeric	1.0451234	2	0.5929995
WAvgRate.perMb	0.4715354	2	0.7899642
ChromTypeX	10.0387615	2	0.0066086

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.

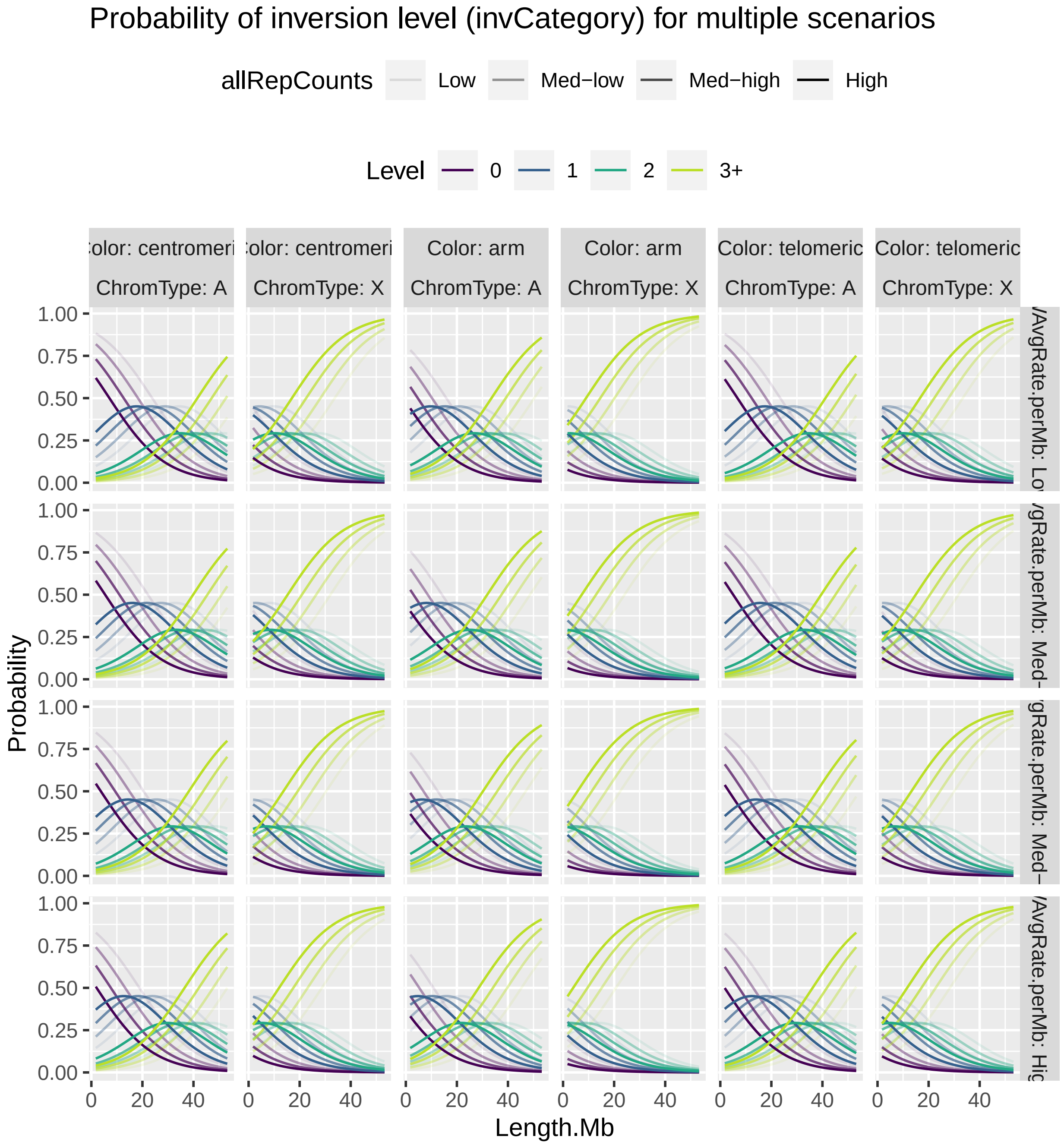


Figure 6: Probabiilty of having 0 to >3 inversions depending on multiple independent variables

NH inversions model

Model fitting

```
## Call:
## polr(formula = myFormula, data = winRegions, Hess = T)
##
## Coefficients:
##                       Value Std. Error t value
## Length.Mb         0.1077147  0.0234081  4.6016
## allRepCounts     -0.0001366  0.0003547 -0.3853
## Colorcentromeric  0.6117393  0.5905533  1.0359
## Colortelomeric   -0.4331394  0.5136583 -0.8432
## WAvgRate.perMb    0.1584862  0.4901615  0.3233
## ChromTypeX       -0.7383865  0.8767738 -0.8422
##
## Intercepts:
##      Value   Std. Error t value
## 0|1   2.7548  1.1339     2.4294
## 1|2   4.7413  1.1970     3.9611
## 2|3+  6.1063  1.2995     4.6990
##
## Residual Deviance: 250.2918
## AIC: 268.2918
```

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

	Value	Std. Error	t value	p value
Length.Mb	0.1077147	0.0234081	4.6016071	0.0000042
allRepCounts	-0.0001366	0.0003547	-0.3852721	0.7000358
Colorcentromeric	0.6117393	0.5905533	1.0358748	0.3002606
Colortelomeric	-0.4331394	0.5136583	-0.8432442	0.3990919
WAvgRate.perMb	0.1584862	0.4901615	0.3233346	0.7464419
ChromTypeX	-0.7383865	0.8767738	-0.8421631	0.3996966
0|1	2.7548434	1.1339460	2.4294308	0.0151226
1|2	4.7413312	1.1969638	3.9611316	0.0000746
2|3+	6.1062851	1.2994974	4.6989591	0.0000026

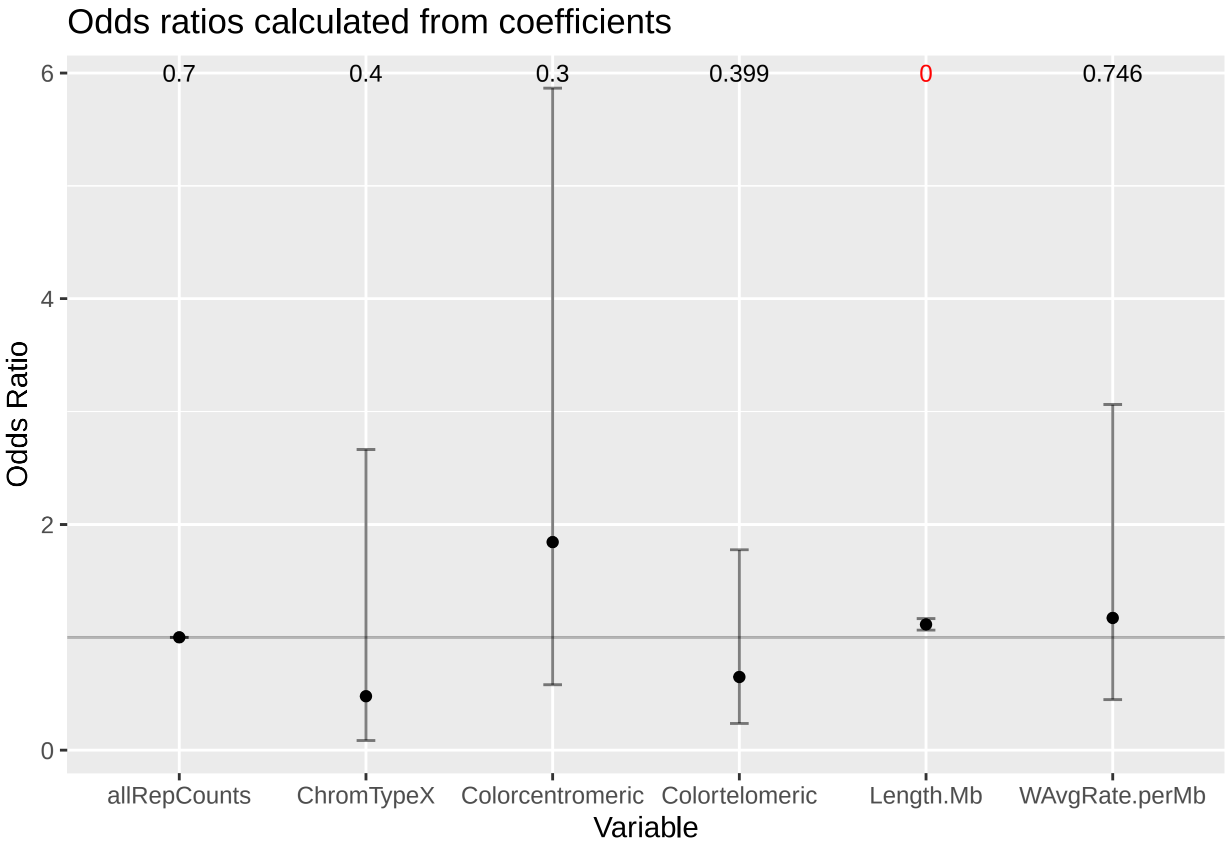
We can also get confidence intervals for the parameter estimates. These can be obtained either by profiling the likelihood function or by using the standard errors and assuming a normal distribution. Note that profiled CIs are not symmetric (although they are usually close to symmetric). If the 95% CI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

	2.5 %	97.5 %
Length.Mb	0.0618357	0.1535936
allRepCounts	-0.0008318	0.0005585
Colorcentromeric	-0.5457240	1.7692025
Colortelomeric	-1.4398911	0.5736124
WAvgRate.perMb	-0.8022128	1.1191852
ChromTypeX	-2.4568315	0.9800584

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	1.1137299	1.0637876	1.166017
allRepCounts	0.9998634	0.9991685	1.000559
Colorcentromeric	1.8436352	0.5794221	5.866173
Colortelomeric	0.6484701	0.2369536	1.774666
WAvgRate.perMb	1.1717357	0.4483358	3.062358
ChromTypeX	0.4778843	0.0857061	2.664612

Example of interpretation: “For 1 unit increase in Length.Mb, a window is 1.1137299 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.

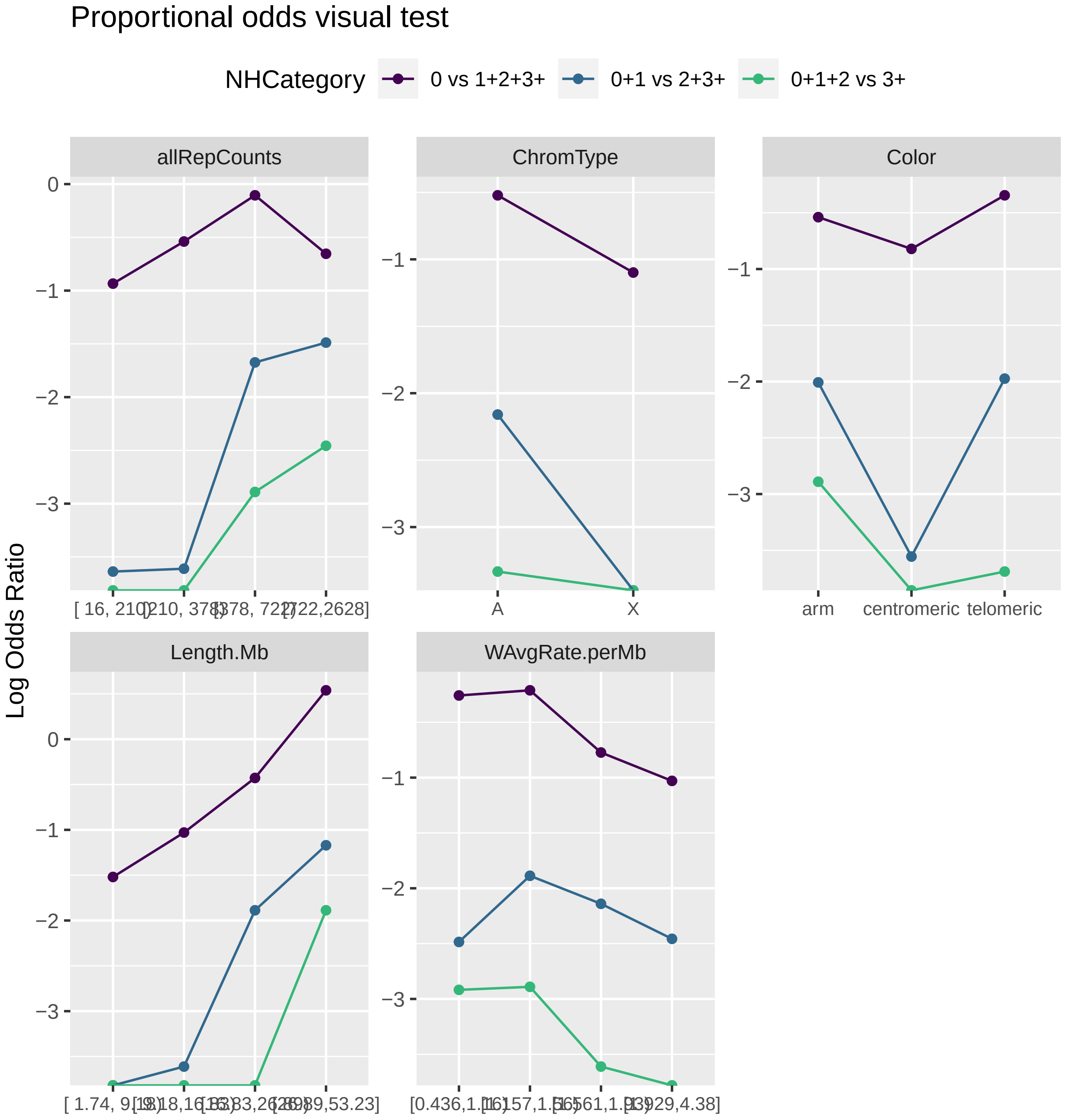
We test the parallel regression assumption with a Brant test:

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

```
## --------------------------------------------------------
## Test for     X2  df  probability
## --------------------------------------------------------
## Omnibus          18.53   12  0.1
## Length.Mb        3.14    2   0.21
## allRepCounts     1.7 2   0.43
## Colorcentromeric 0.01    2   1
## Colortelomeric       4.22    2   0.12
## WAvgRate.perMb       9.95    2   0.01
## ChromTypeX       0   2   1
## --------------------------------------------------------
##
## H0: Parallel Regression Assumption holds
```

	X2	df	probability
Omnibus	18.5271510	12	0.1005967
Length.Mb	3.1373185	2	0.2083243
allRepCounts	1.7033597	2	0.4266975
Colorcentromeric	0.0068235	2	0.9965940
Colortelomeric	4.2180941	2	0.1213536
WAvgRate.perMb	9.9454172	2	0.0069244
ChromTypeX	0.0001147	2	0.9999426

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.

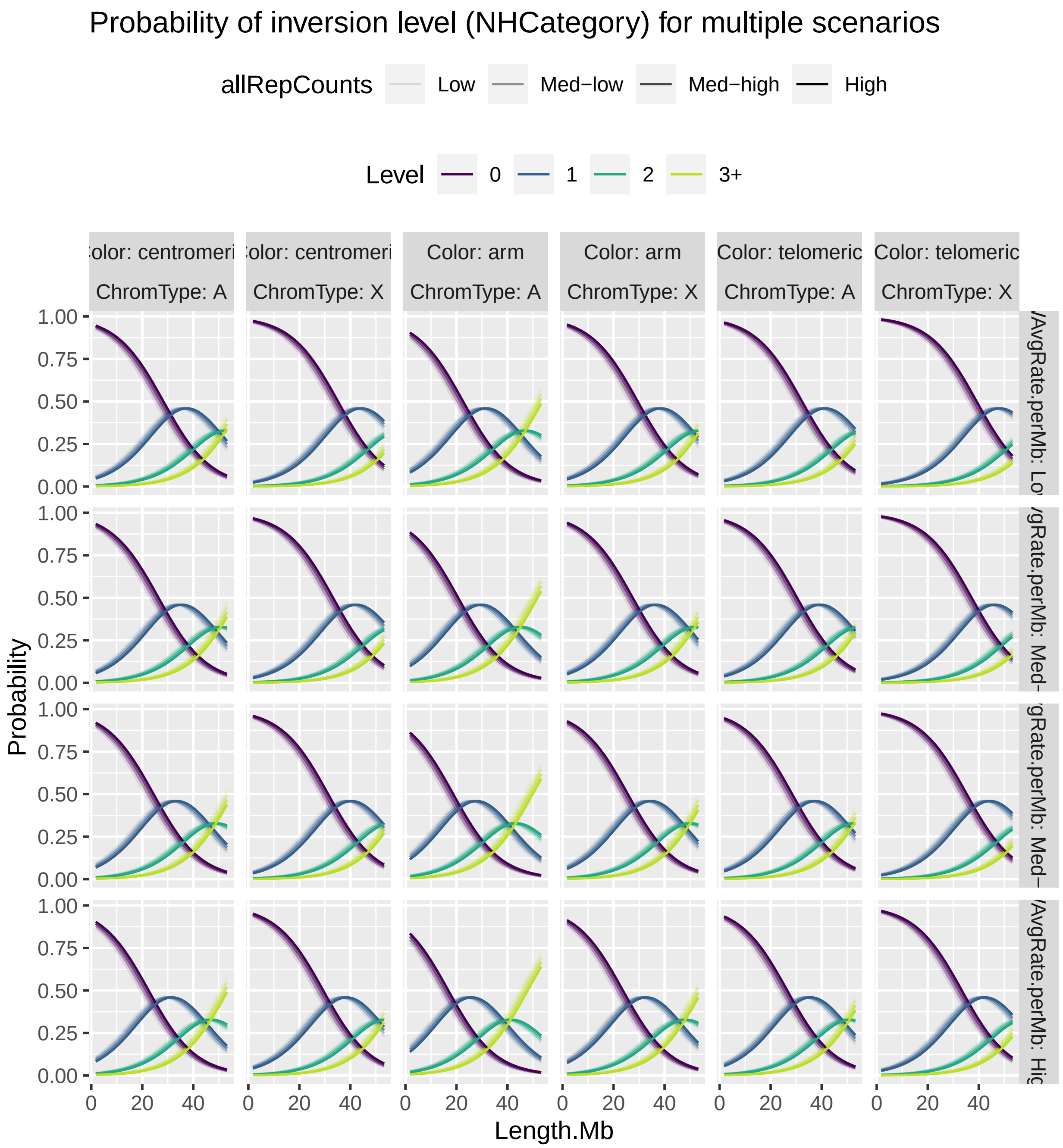


Figure 7: Probabiilty of having 0 to >3 inversions depending on multiple independent variables

NAHR inversions model

Model fitting

```
## Call:
## polr(formula = myFormula, data = winRegions, Hess = T)
##
## Coefficients:
##                     Value Std. Error t value
## Length.Mb        0.026554  0.0236511 1.12273
## allRepCounts     0.001089  0.0003749 2.90569
## Colorcentromeric 0.535665  0.6897193 0.77664
## Colortelomeric   0.546808  0.5555471 0.98427
## WAvgRate.perMb   0.022650  0.5537166 0.04091
## ChromTypeX       3.237525  0.8626995 3.75278
##
## Intercepts:
##      Value  Std. Error t value
## 0|1  2.7337 1.2798     2.1360
## 1|2  4.9112 1.3663     3.5945
## 2|3+ 6.2098 1.4733     4.2149
##
## Residual Deviance: 195.3175
## AIC: 213.3175
```

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

	Value	Std. Error	t value	p value
Length.Mb	0.0265537	0.0236511	1.1227273	0.2615533
allRepCounts	0.0010894	0.0003749	2.9056899	0.0036644
Colorcentromeric	0.5356653	0.6897193	0.7766424	0.4373698
Colortelomeric	0.5468079	0.5555471	0.9842692	0.3249832
WAvgRate.perMb	0.0226504	0.5537166	0.0409061	0.9673708
ChromTypeX	3.2375252	0.8626995	3.7527842	0.0001749
0|1	2.7337203	1.2798038	2.1360464	0.0326756
1|2	4.9112146	1.3663127	3.5945027	0.0003250
2|3+	6.2097957	1.4733015	4.2148844	0.0000250

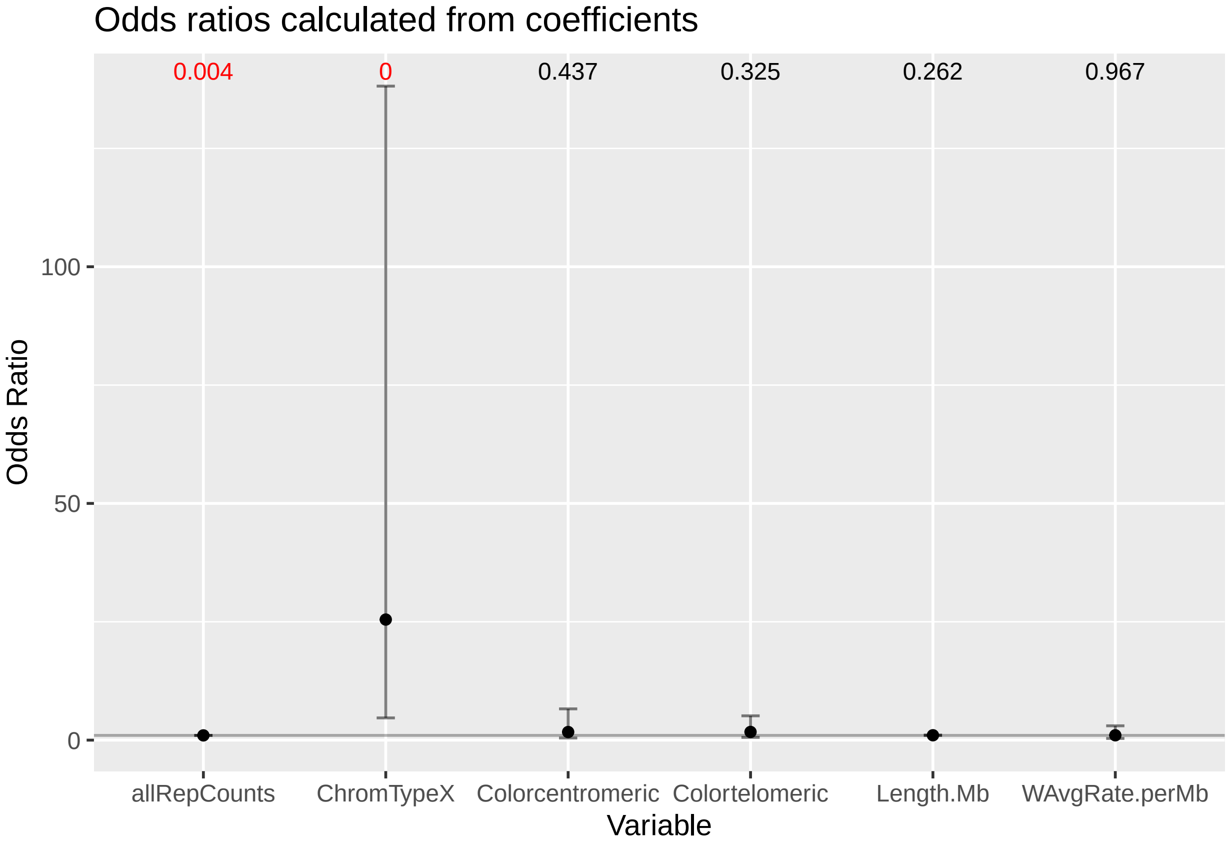
We can also get confidence intervals for the parameter estimates. These can be obtained either by profiling the likelihood function or by using the standard errors and assuming a normal distribution. Note that profiled CIs are not symmetric (although they are usually close to symmetric). If the 95% CI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

	2.5 %	97.5 %
Length.Mb	-0.0198015	0.0729089
allRepCounts	0.0003546	0.0018242
Colorcentromeric	-0.8161597	1.8874902
Colortelomeric	-0.5420444	1.6356601
WAvgRate.perMb	-1.0626141	1.1079149
ChromTypeX	1.5466652	4.9283852

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	1.026909	0.9803932	1.075633
allRepCounts	1.001090	1.0003546	1.001826
Colorcentromeric	1.708584	0.4421263	6.602776
Colortelomeric	1.727729	0.5815581	5.132845
WAvgRate.perMb	1.022909	0.3455513	3.028038
ChromTypeX	25.470609	4.6957844	138.156240

Example of interpretation: "For 1 unit increase in Length.Mb, a window is 1.0269094 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.

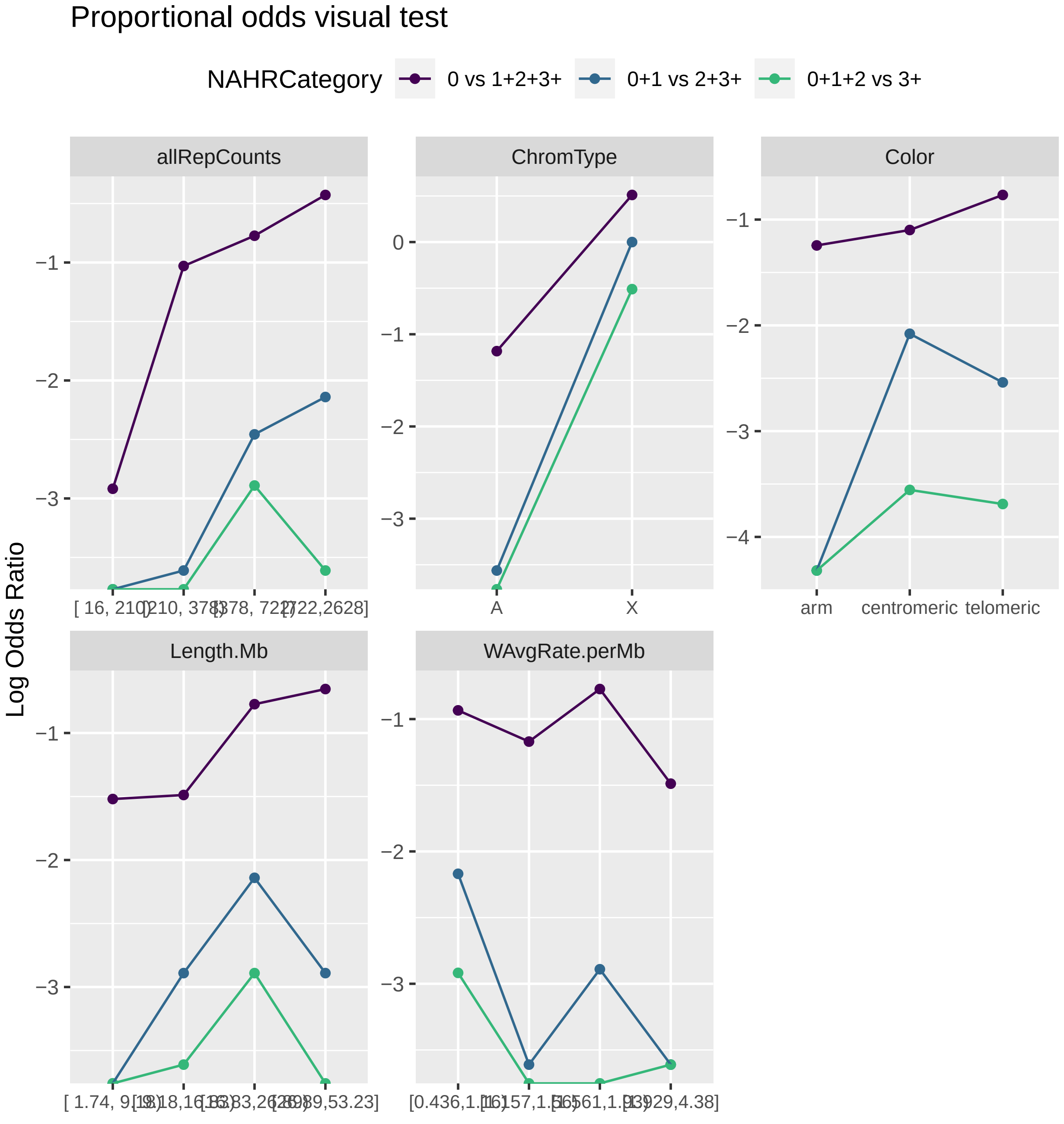
We test the parallel regression assumption with a Brant test:

```
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	18.5271510	12	0.1005967
Length.Mb	3.1373185	2	0.2083243
allRepCounts	1.7033597	2	0.4266975
Colorcentromeric	0.0068235	2	0.9965940
Colortelomeric	4.2180941	2	0.1213536
WAvgRate.perMb	9.9454172	2	0.0069244
ChromTypeX	0.0001147	2	0.9999426

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.

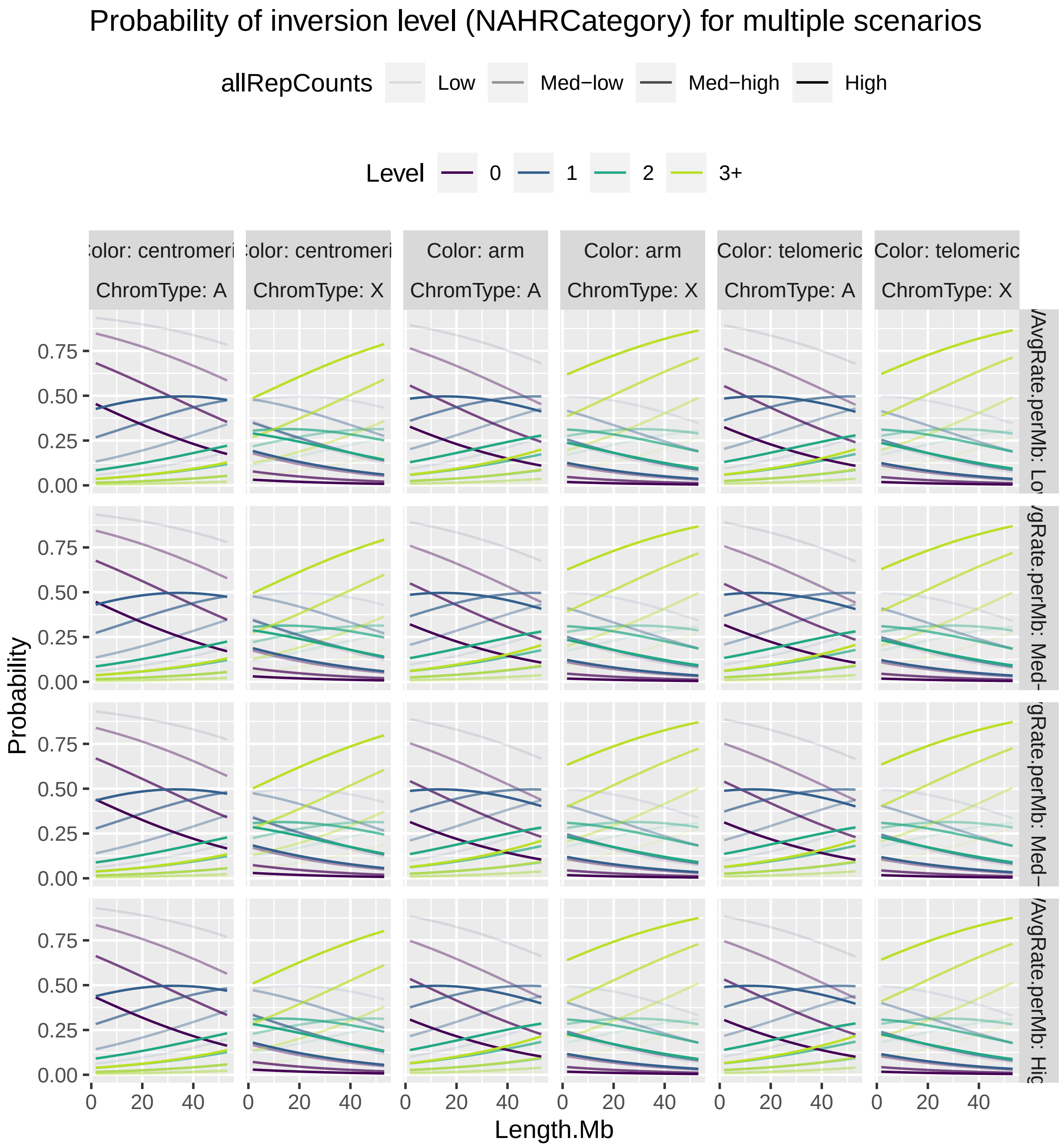


Figure 8: Probabiilty of having 0 to >3 inversions depending on multiple independent variables

Scaled variables

Total inversions model

Model fitting

```
## Call:
## polr(formula = myFormula, data = winRegions, Hess = T)
##
## Coefficients:
##                         Value Std. Error t value
## Length.Mb.Scaled      1.02171     0.2331 4.38271
## allRepCounts.Scaled   0.32303     0.1592 2.02876
## Colorcentromeric      0.73020     0.5432 1.34430
## Colortelomeric        0.03203     0.4484 0.07143
## WAvgRate.perMb.Scaled 0.07198     0.2567 0.28035
## ChromTypeX            2.24615     0.7884 2.84909
##
## Intercepts:
##      Value  Std. Error t value
## 0|1  0.0363 0.2455     0.1477
## 1|2  1.9822 0.2985     6.6406
## 2|3+ 3.1882 0.3994     7.9829
##
## Residual Deviance: 311.8409
## AIC: 329.8409
```

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	1.0217131	0.2331239	4.3827051	0.0000117
allRepCounts.Scaled	0.3230349	0.1592279	2.0287581	0.0424829
Colorcentromeric	0.7301970	0.5431817	1.3442960	0.1788527
Colortelomeric	0.0320281	0.4483549	0.0714346	0.9430518
WAvgRate.perMb.Scaled	0.0719765	0.2567394	0.2803485	0.7792101
ChromTypeX	2.2461527	0.7883746	2.8490932	0.0043844
0|1	0.0362522	0.2455276	0.1476502	0.8826188
1|2	1.9822454	0.2985042	6.6405942	0.0000000
2|3+	3.1882180	0.3993807	7.9829039	0.0000000

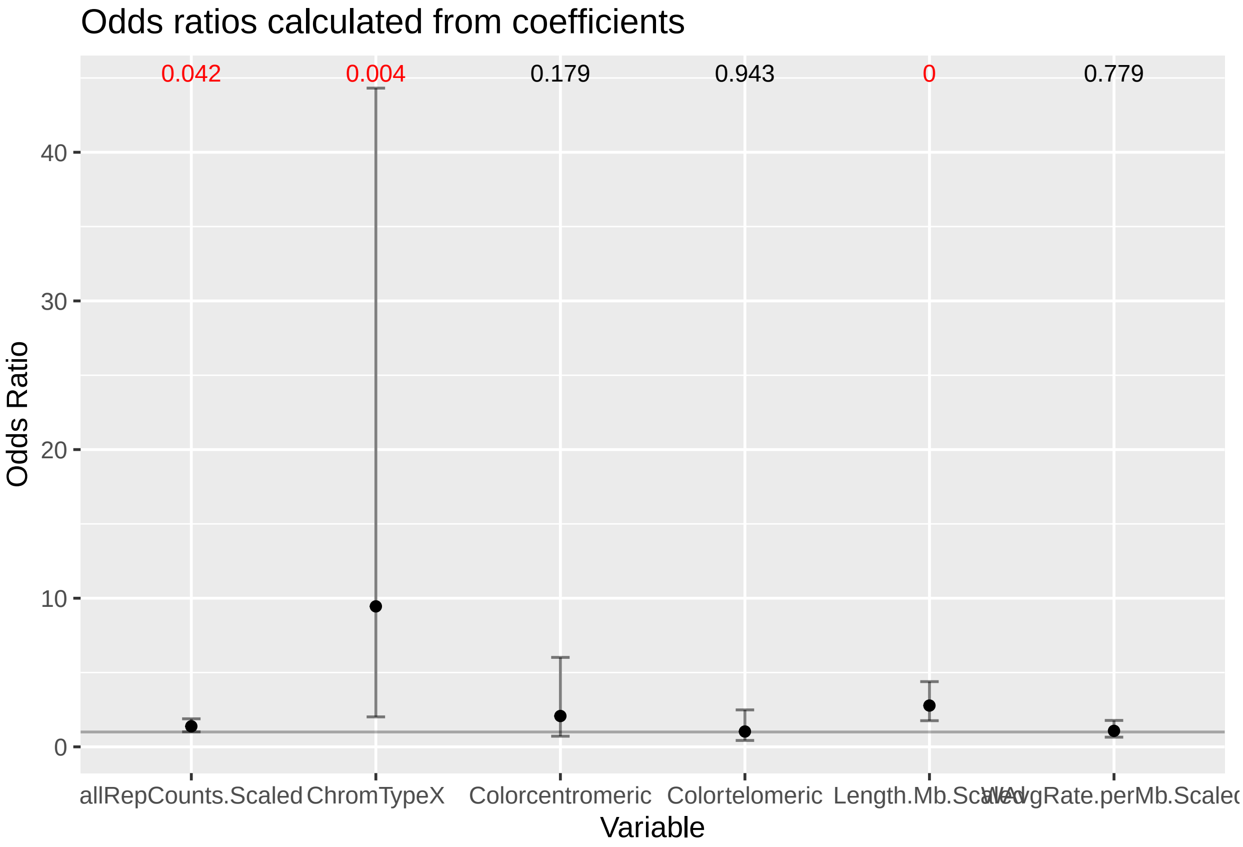
We can also get confidence intervals for the parameter estimates. These can be obtained either by profiling the likelihood function or by using the standard errors and assuming a normal distribution. Note that profiled CIs are not symmetric (although they are usually close to symmetric). If the 95% CI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

	2.5 %	97.5 %
Length.Mb.Scaled	0.5647987	1.4786275
allRepCounts.Scaled	0.0109539	0.6351159
Colorcentromeric	-0.3344196	1.7948135
Colortelomeric	-0.8467315	0.9107876
WAvgRate.perMb.Scaled	-0.4312235	0.5751765
ChromTypeX	0.7009669	3.7913385

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	2.777950	1.7590937	4.386920
allRepCounts.Scaled	1.381314	1.0110142	1.887241
Colorcentromeric	2.075489	0.7157534	6.018352
Colortelomeric	1.032547	0.4288142	2.486280
WAvgRate.perMb.Scaled	1.074630	0.6497137	1.777444
ChromTypeX	9.451304	2.0157007	44.315677

Example of interpretation: "For 1 unit increase in Length.Mb.Scaled, a window is 2.7779496 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.

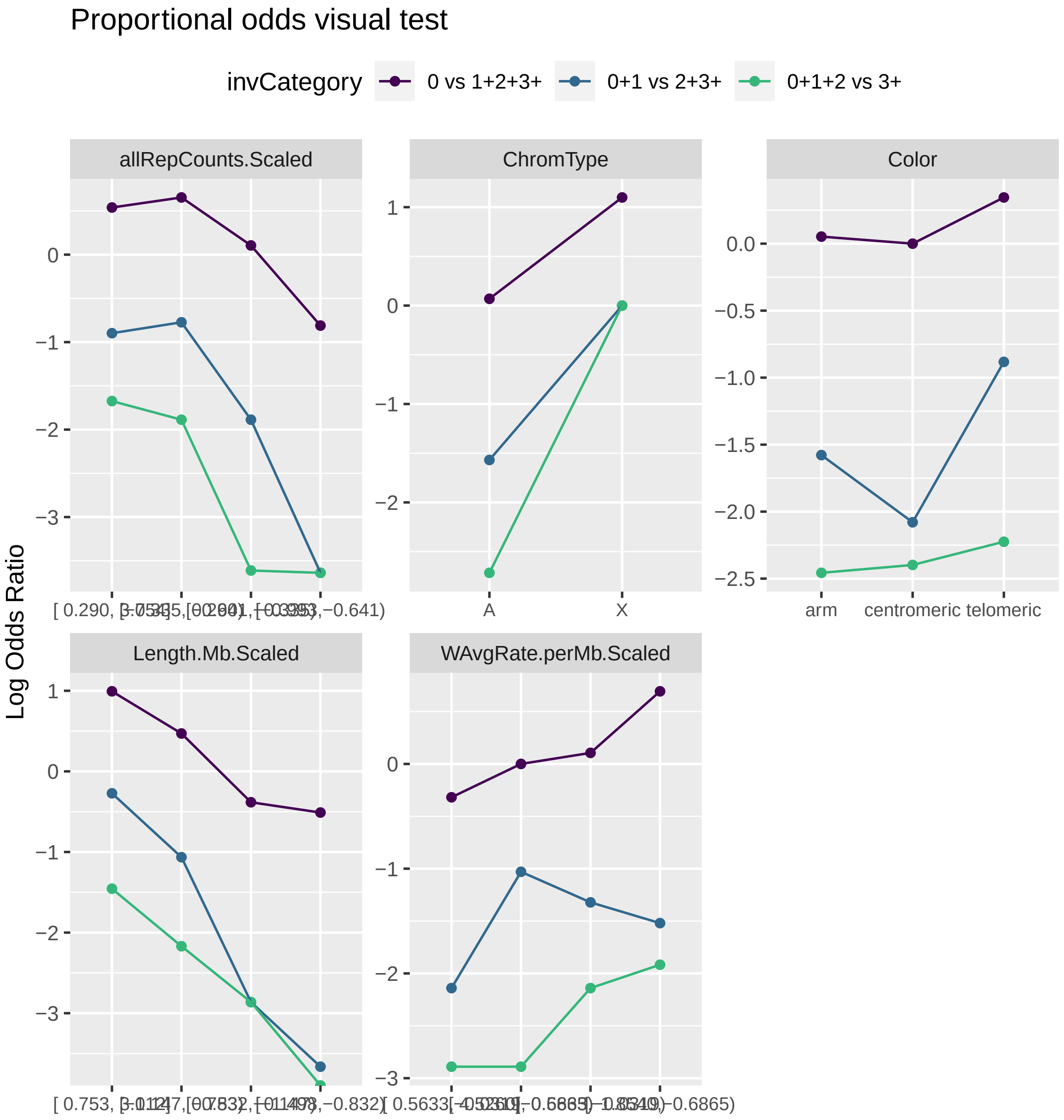
We test the parallel regression assumption with a Brant test:

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

```
## ----------------------------------------------------
## Test for     X2  df  probability
## ----------------------------------------------------
## Omnibus          42.36   12  0
## Length.Mb.Scaled 5.32    2   0.07
## allRepCounts.Scaled  0.1 2   0.95
## Colorcentromeric 1.71    2   0.43
## Colortelomeric       1.05    2   0.59
## WAvgRate.perMb.Scaled    0.47    2   0.79
## ChromTypeX       10.04   2   0.01
## ----------------------------------------------------
##
## H0: Parallel Regression Assumption holds
```

	X2	df	probability
Omnibus	42.3581930	12	0.0000290
Length.Mb.Scaled	5.3215725	2	0.0698932
allRepCounts.Scaled	0.0978698	2	0.9522431
Colorcentromeric	1.7067240	2	0.4259804
Colortelomeric	1.0451234	2	0.5929995
WAvgRate.perMb.Scaled	0.4715354	2	0.7899642
ChromTypeX	10.0387615	2	0.0066086

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.

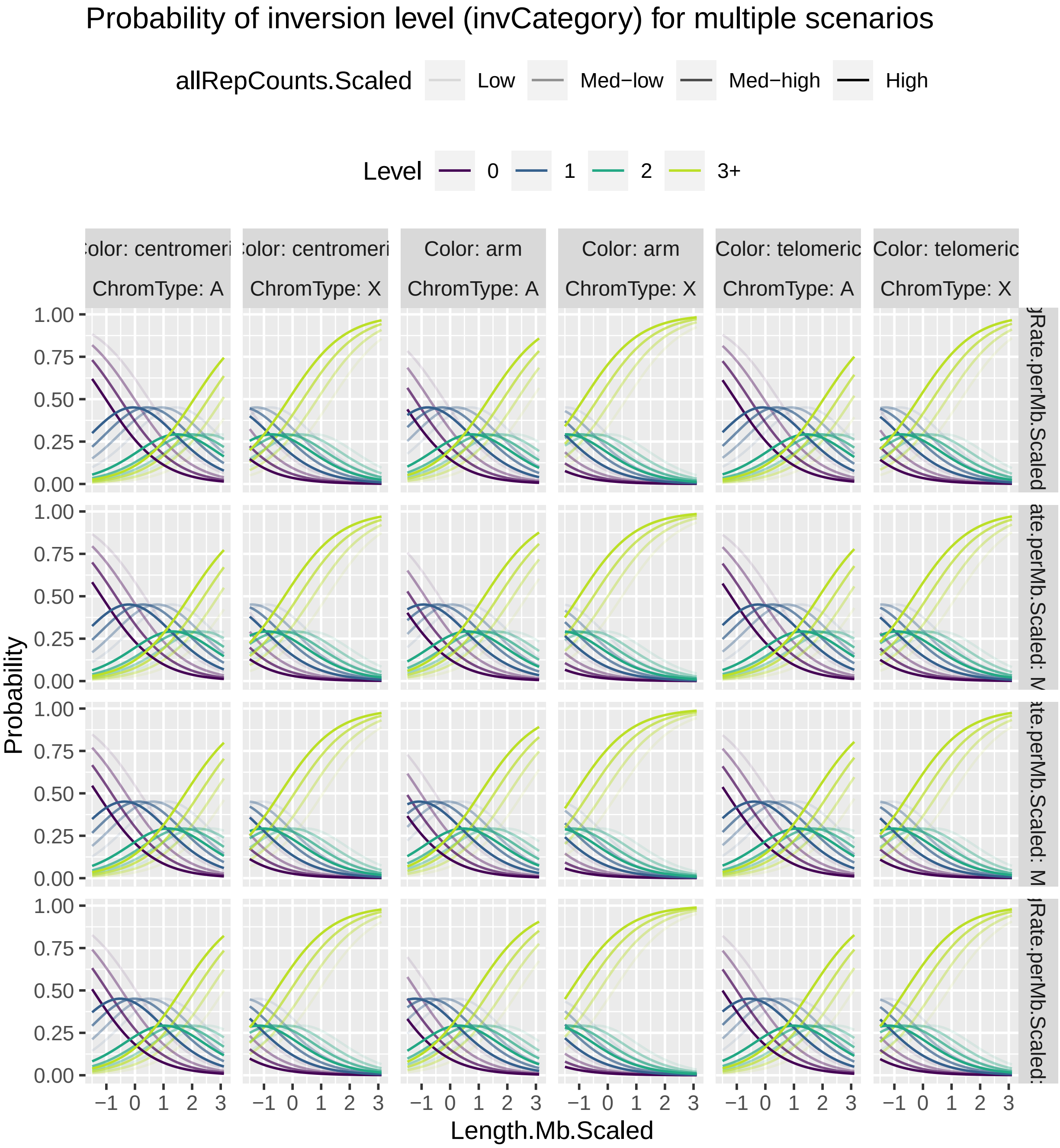


Figure 9: Probabiilty of having 0 to >3 inversions depending on multiple independent variables

NH inversions model

Model fitting

```
## Call:
## polr(formula = myFormula, data = winRegions, Hess = T)
##
## Coefficients:
##                           Value Std. Error t value
## Length.Mb.Scaled        1.20312     0.2621  4.5908
## allRepCounts.Scaled    -0.07520     0.1878 -0.4004
## Colorcentromeric        0.61175     0.5906  1.0357
## Colortelomeric         -0.43311     0.5137 -0.8432
## WAvgRate.perMb.Scaled  0.09786     0.3027  0.3233
## ChromTypeX             -0.73836     0.8768 -0.8421
##
## Intercepts:
##      Value   Std. Error t value
## 0|1   0.5914  0.2646     2.2346
## 1|2   2.5779  0.3604     7.1522
## 2|3+  3.9428  0.5360     7.3564
##
## Residual Deviance: 250.2918
## AIC: 268.2918
```

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	1.2031248	0.2620727	4.5908058	0.0000044
allRepCounts.Scaled	-0.0751950	0.1877772	-0.4004481	0.6888265
Colorcentromeric	0.6117476	0.5906442	1.0357296	0.3003283
Colortelomeric	-0.4331068	0.5136590	-0.8431796	0.3991280
WAvgRate.perMb.Scaled	0.0978564	0.3026726	0.3233079	0.7464621
ChromTypeX	-0.7383584	0.8768492	-0.8420587	0.3997551
0|1	0.5913824	0.2646481	2.2345986	0.0254437
1|2	2.5779023	0.3604341	7.1522157	0.0000000
2|3+	3.9428418	0.5359749	7.3563920	0.0000000

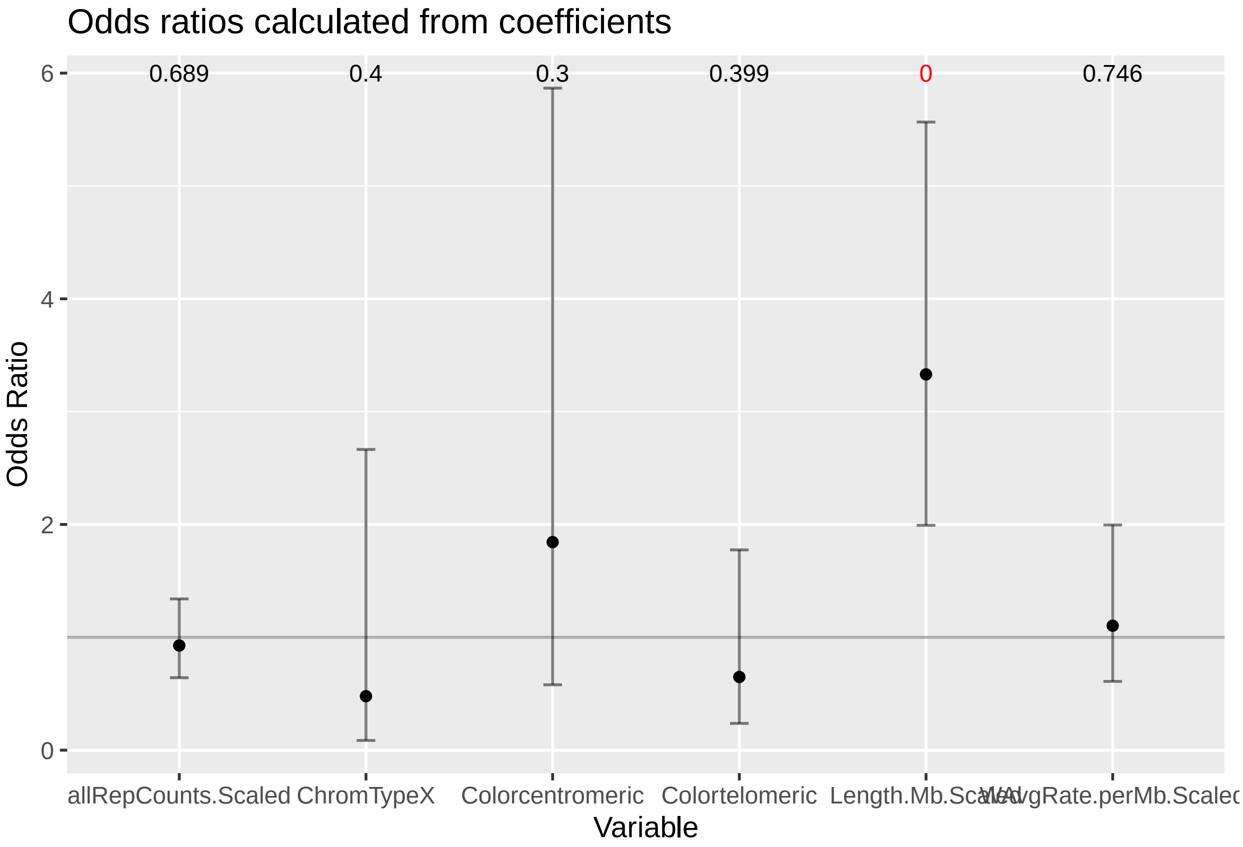
We can also get confidence intervals for the parameter estimates. These can be obtained either by profiling the likelihood function or by using the standard errors and assuming a normal distribution. Note that profiled CIs are not symmetric (although they are usually close to symmetric). If the 95% CI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

	2.5 %	97.5 %
Length.Mb.Scaled	0.6894718	1.7167778
allRepCounts.Scaled	-0.4432315	0.2928415
Colorcentromeric	-0.5458937	1.7693890
Colortelomeric	-1.4398599	0.5736463
WAvgRate.perMb.Scaled	-0.4953710	0.6910838
ChromTypeX	-2.4569512	0.9802343

We convert the coefficients into odds ratios. To get the OR and confidence intervals, we just exponentiate the estimates and confidence intervals (here I used the likelihood confidence intervals).

	Odds Ratio	2.5%	97.5%
Length.Mb.Scaled	3.3305078	1.9926627	5.566563
allRepCounts.Scaled	0.9275626	0.6419586	1.340230
Colorcentromeric	1.8436506	0.5793238	5.867267
Colortelomeric	0.6484912	0.2369610	1.774727
WAvgRate.perMb.Scaled	1.1028045	0.6093448	1.995878
ChromTypeX	0.4778978	0.0856958	2.665081

Example of interpretation: "For 1 unit increase in Length.Mb.Scaled, a window is 3.3305078 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.

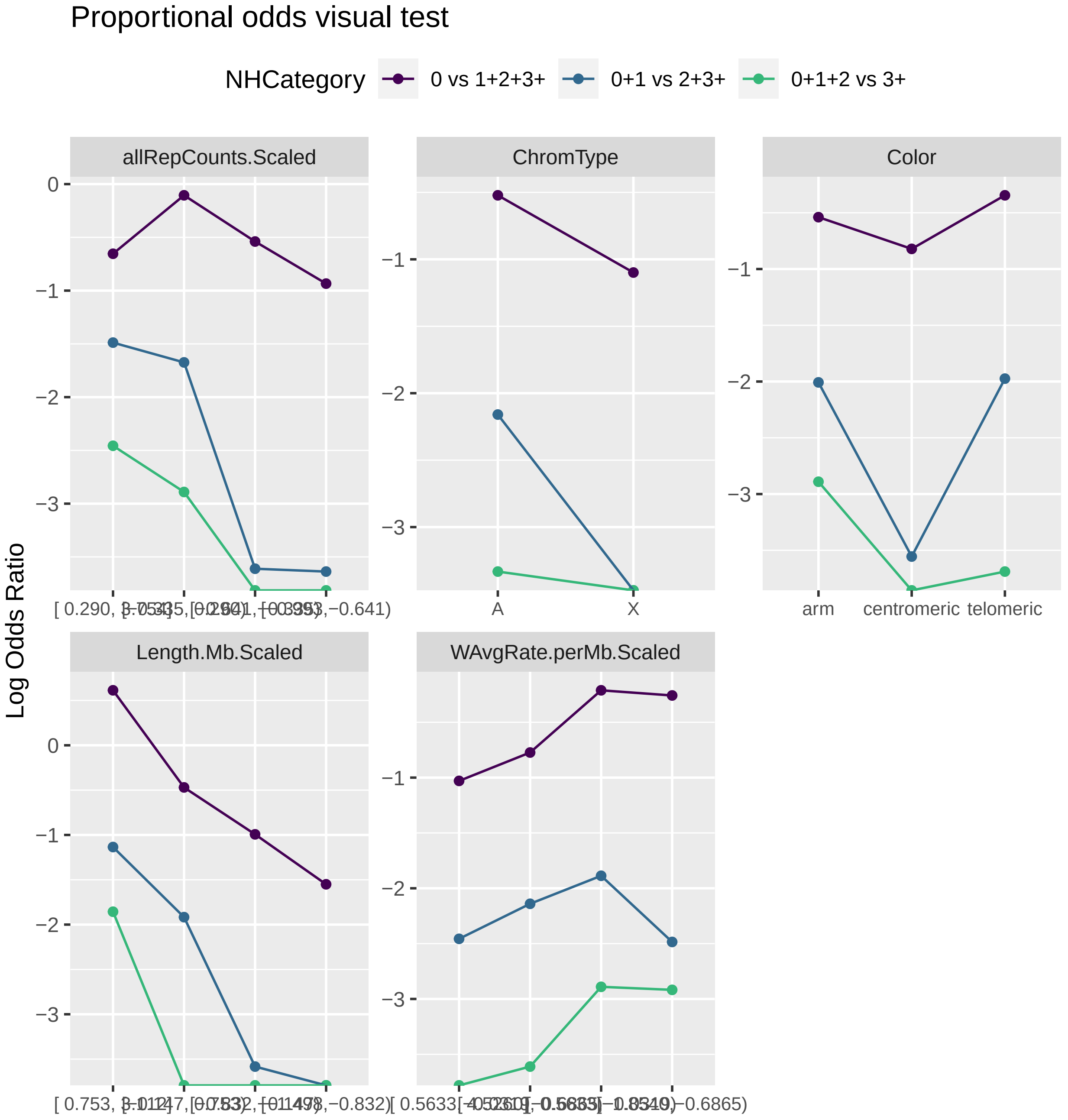
We test the parallel regression assumption with a Brant test:

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

```
## --------------------------------------------------------
## Test for     X2  df  probability
## --------------------------------------------------------
## Omnibus          18.53   12  0.1
## Length.Mb.Scaled 3.14    2   0.21
## allRepCounts.Scaled  1.7 2   0.43
## Colorcentromeric 0.01    2   1
## Colortelomeric       4.22    2   0.12
## WAvgRate.perMb.Scaled    9.95    2   0.01
## ChromTypeX       0   2   1
## --------------------------------------------------------
##
## H0: Parallel Regression Assumption holds
```

	X2	df	probability
Omnibus	18.5271510	12	0.1005967
Length.Mb.Scaled	3.1373185	2	0.2083243
allRepCounts.Scaled	1.7033597	2	0.4266975
Colorcentromeric	0.0068235	2	0.9965940
Colortelomeric	4.2180941	2	0.1213536
WAvgRate.perMb.Scaled	9.9454172	2	0.0069244
ChromTypeX	0.0001147	2	0.9999426

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.

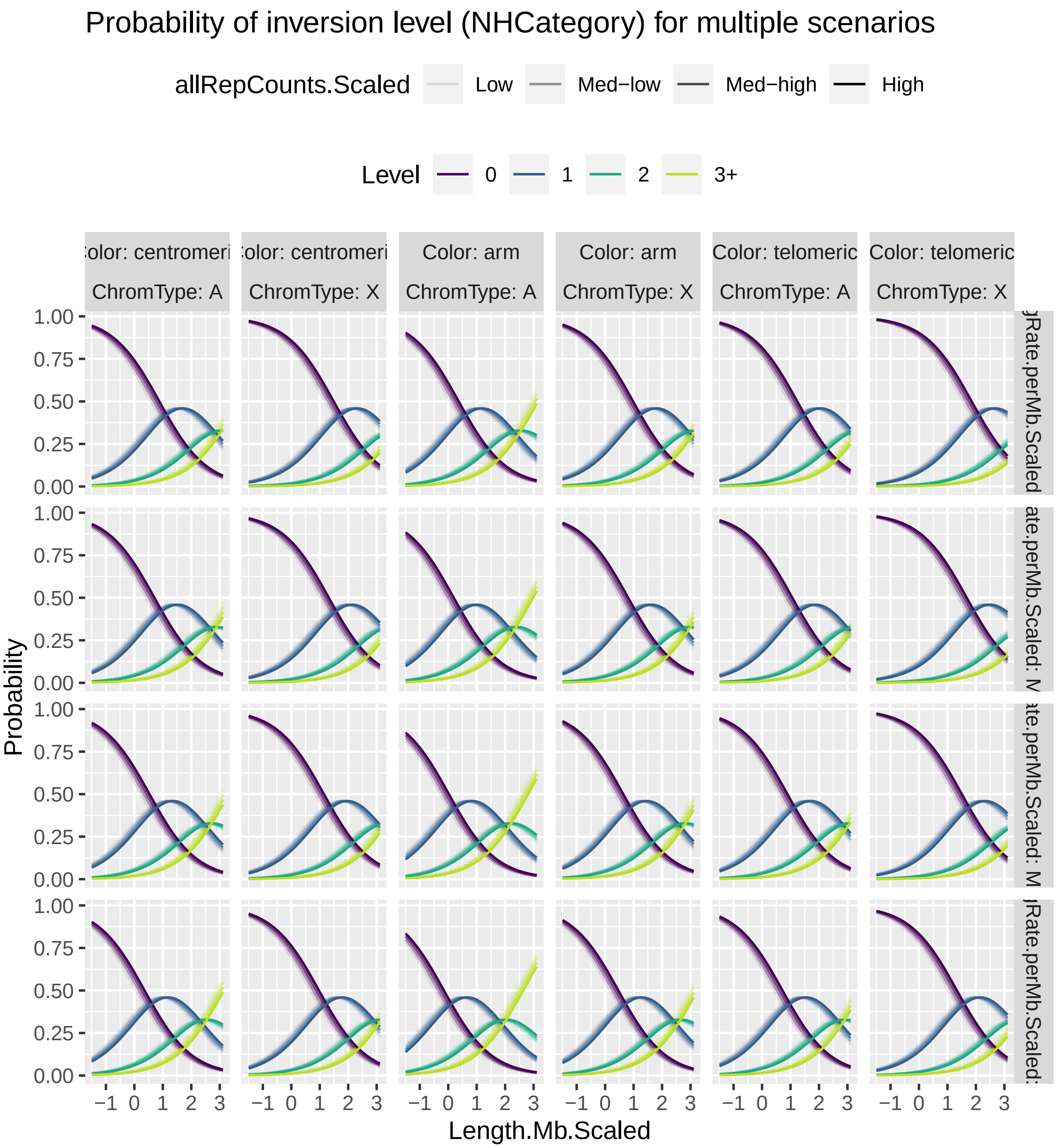


Figure 10: Probabiilty of having 0 to >3 inversions depending on multiple independent variables

NAHR inversions model

Model fitting

```
## Call:
## polr(formula = myFormula, data = winRegions, Hess = T)
##
## Coefficients:
##                        Value Std. Error t value
## Length.Mb.Scaled      0.29658     0.2635 1.12538
## allRepCounts.Scaled   0.59944     0.1912 3.13476
## Colorcentromeric      0.53562     0.6894 0.77698
## Colortelomeric        0.54681     0.5553 0.98479
## WAvgRate.perMb.Scaled 0.01398     0.3423 0.04084
## ChromTypeX            3.23745     0.8611 3.75957
##
## Intercepts:
##      Value  Std. Error t value
## 0|1  1.5945 0.3101     5.1412
## 1|2  3.7720 0.5012     7.5267
## 2|3+ 5.0706 0.7425     6.8286
##
## Residual Deviance: 195.3175
## AIC: 213.3175
```

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	0.2965768	0.2635338	1.1253840	0.2604263
allRepCounts.Scaled	0.5994429	0.1912245	3.1347598	0.0017200
Colorcentromeric	0.5356218	0.6893679	0.7769752	0.4371734
Colortelomeric	0.5468083	0.5552550	0.9847876	0.3247284
WAvgRate.perMb.Scaled	0.0139796	0.3422896	0.0408413	0.9674224
ChromTypeX	3.2374537	0.8611240	3.7595674	0.0001702
0|1	1.5945335	0.3101455	5.1412433	0.0000003
1|2	3.7719992	0.5011507	7.5266766	0.0000000
2|3+	5.0705831	0.7425463	6.8286422	0.0000000

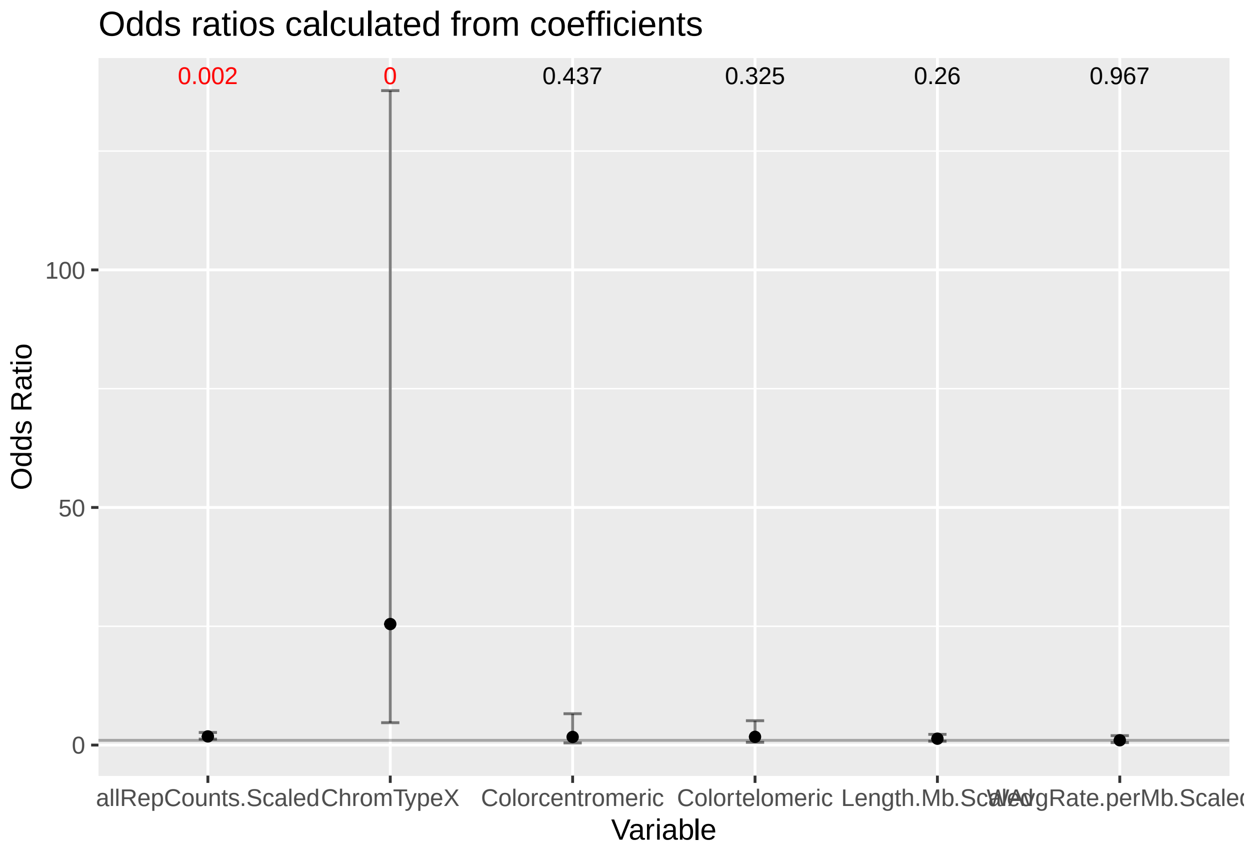
We can also get confidence intervals for the parameter estimates. These can be obtained either by profiling the likelihood function or by using the standard errors and assuming a normal distribution. Note that profiled CIs are not symmetric (although they are usually close to symmetric). If the 95% CI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

	2.5 %	97.5 %
Length.Mb.Scaled	-0.2199401	0.8130936
allRepCounts.Scaled	0.2246498	0.9742361
Colorcentromeric	-0.8155145	1.8867580
Colortelomeric	-0.5414716	1.6350882
WAvgRate.perMb.Scaled	-0.6568958	0.6848549
ChromTypeX	1.5496817	4.9252257

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	1.345246	0.8025669	2.254873
allRepCounts.Scaled	1.821104	1.2518842	2.649143
Colorcentromeric	1.708510	0.4424116	6.597944
Colortelomeric	1.727730	0.5818913	5.129910
WAvgRate.perMb.Scaled	1.014078	0.5184583	1.983484
ChromTypeX	25.468788	4.7099707	137.720422

Example of interpretation: "For 1 unit increase in Length.Mb.Scaled, a window is 1.3452458 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.

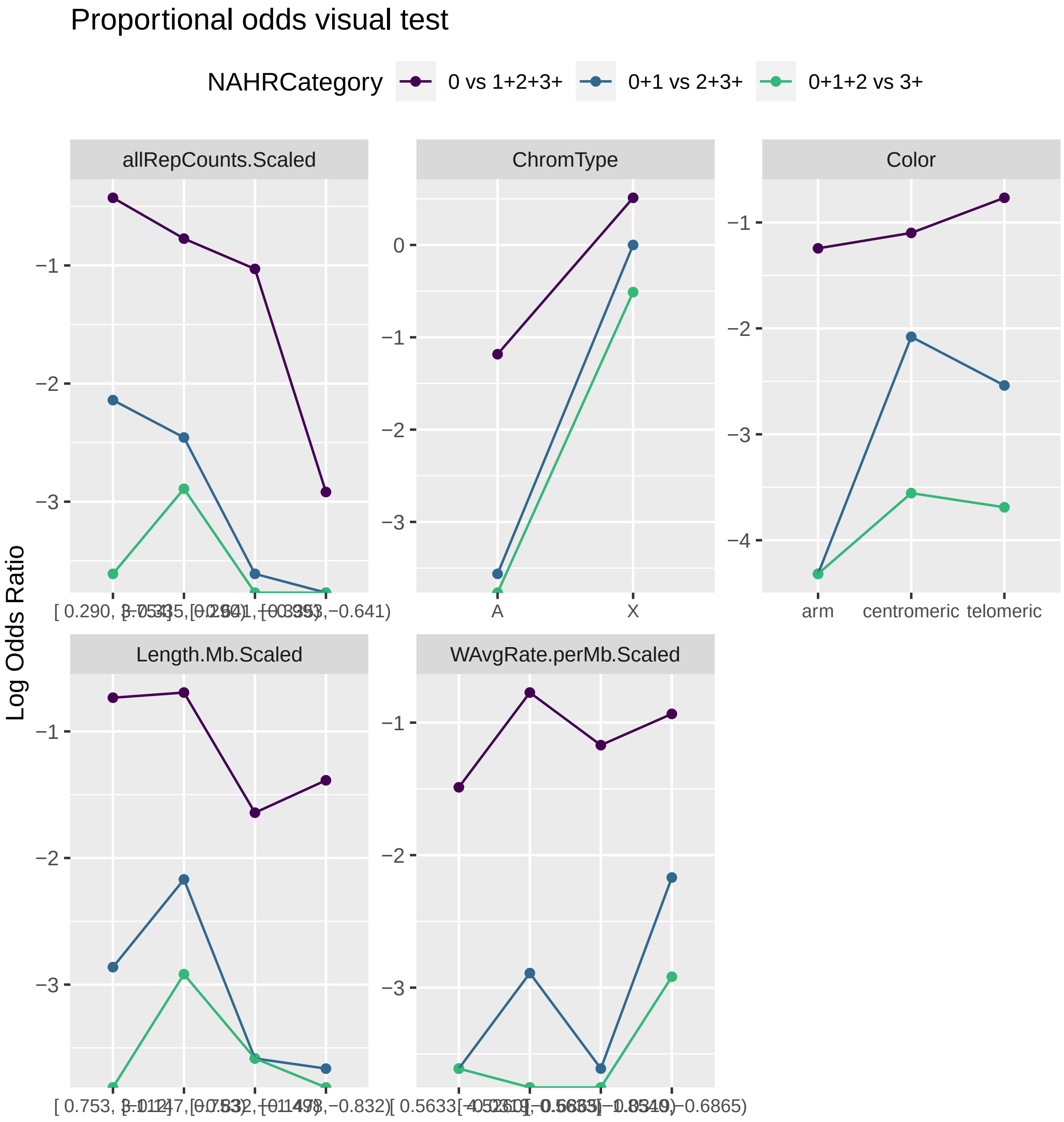
We test the parallel regression assumption with a Brant test:

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

```
## ----------------------------------------------------
## Test for     X2  df  probability
## ----------------------------------------------------
## Omnibus          -0.67   12  1
## 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	-0.6744064	12	1.000000
Length.Mb.Scaled	0.0000000	2	1.000000
allRepCounts.Scaled	0.0000020	2	0.999999
Colorcentromeric	0.0000000	2	1.000000
Colortelomeric	0.0000001	2	1.000000
WAvgRate.perMb.Scaled	0.0000000	2	1.000000
ChromTypeX	-0.0000244	2	1.000000

We can also evaluate the parallel regression visually. We transform the ordinal dependent variable with k categories into a series of k-1 binary variables that indicate whether the dependent value is above or below a cutpoint (e.g. windows with at least 2 inversions vs windows with less than 2 inversions). We then calculate the observed Log Odds Ratio for each binary variable across multiple value ranges of the independent variables. The lines should be approximately parallel, that each independent variable affects the probability of increasing by 1 level the inversion count in the same way, for all transitions, and that we don't need a specific model for each level increase.



Predicted probabilites

Although our objective is to describe the dataset, predicted probabilities are usually easier to understand than either the coefficients or the Odds Ratios.

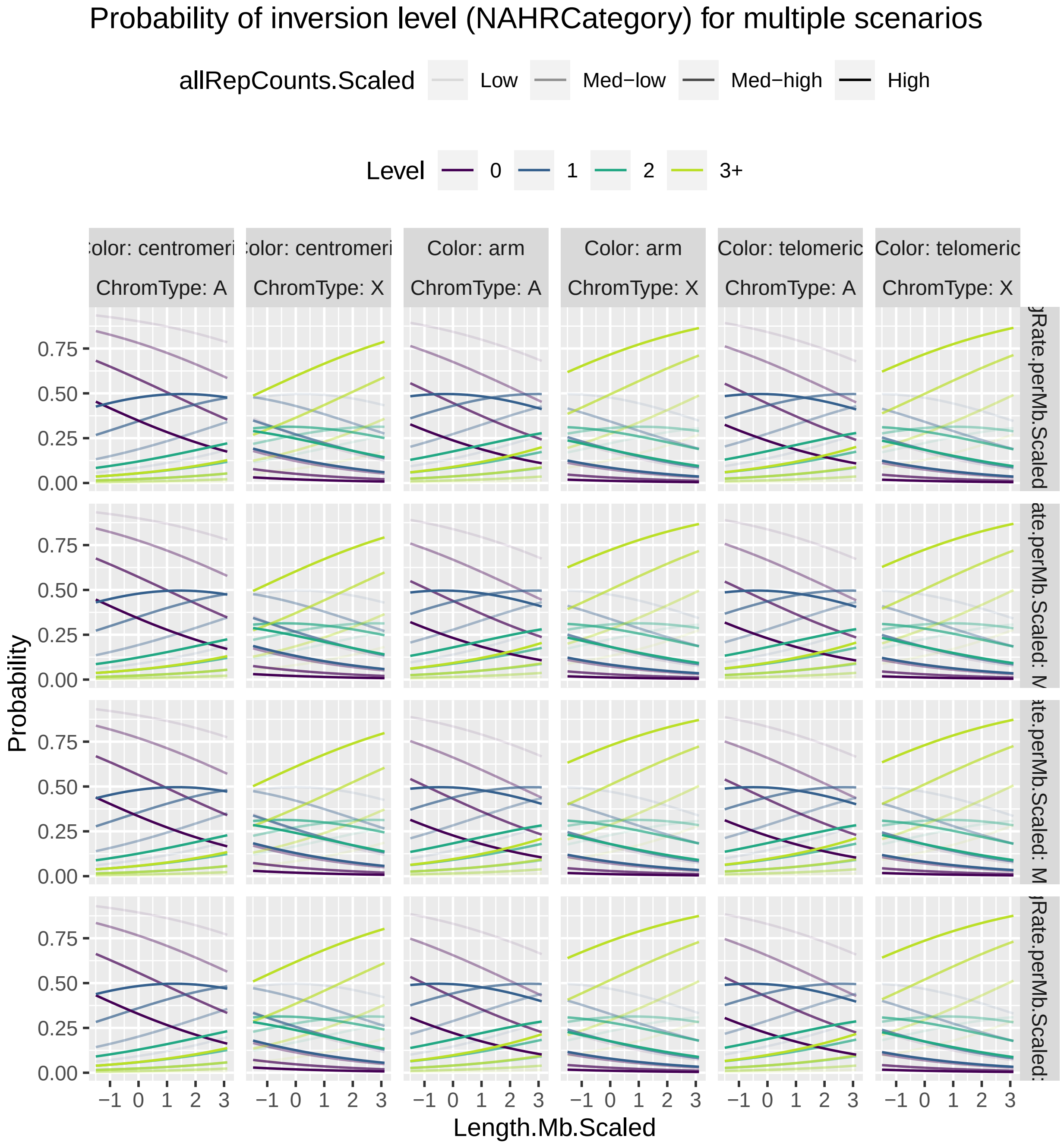


Figure 11: Probabiilty of having 0 to >3 inversions depending on multiple independent variables

Descriptive categories

Descriptive statistics

Raw data:

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

For each window, I calculated the number of total inversions, NH inversions, and NAHR inversions, the window length in Mb, number of repeats and the average recombination rate in cM/Mb.

I want to perform Ordinal Logistic Regressions on different subsets of the data. The assumptions of the Ordinal Logistic Regression are as follow:

1. The dependent variable is ordered.
2. One or more of the independent variables are either continuous, categorical or ordinal.
3. No multi-collinearity.
4. Proportional odds.

I show the data distributions in the figure below. The inversion counts have only a number of possible options, so they can be considered an ordinal variable. The independent variables are continuous and categorical, so assumptions 1 and 2 are satisfied

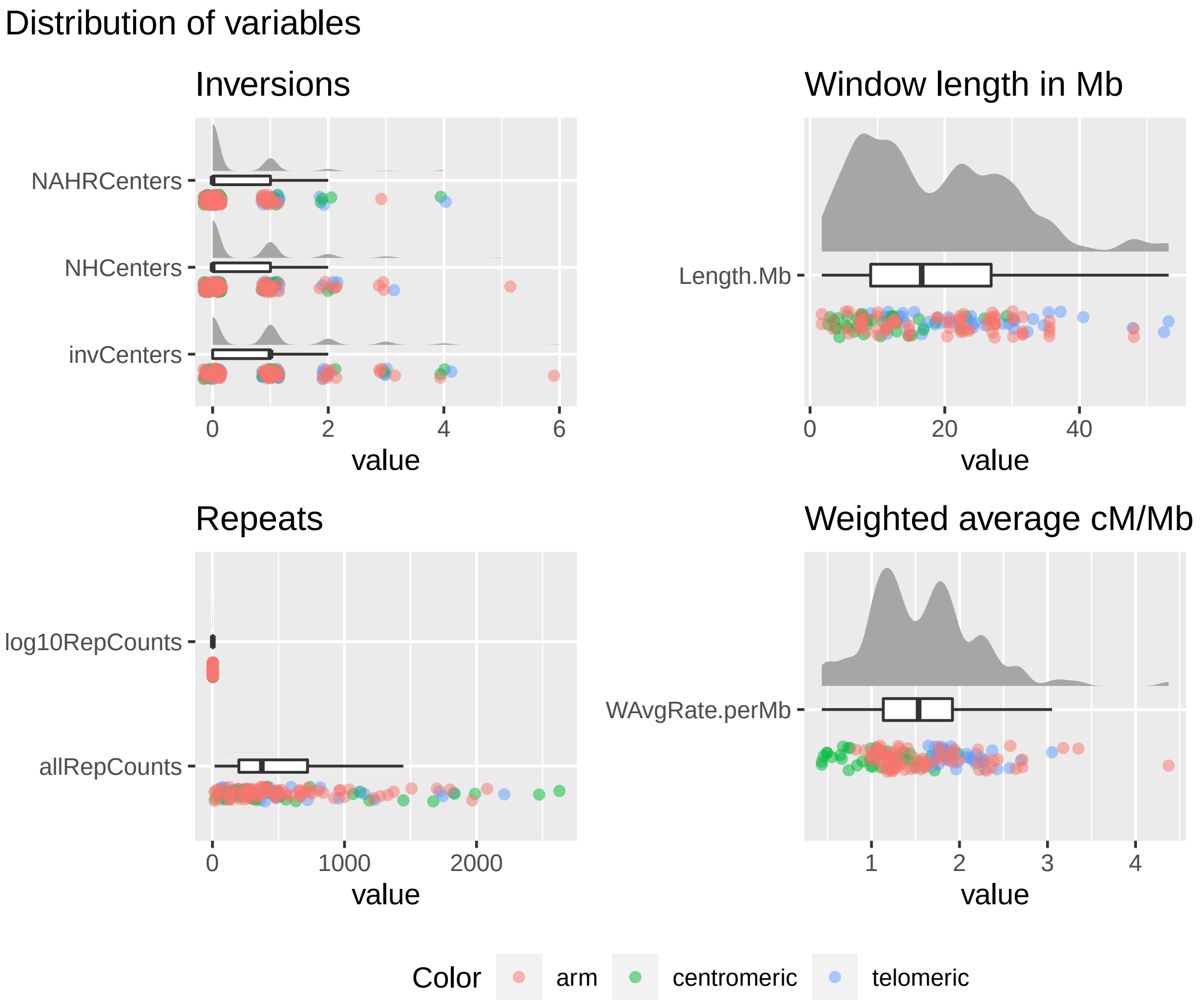


Figure 12: Distribution of variables.

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

Table 32: Original counts

CountGroups	invCenters	NHCenters	NAHRCenters
0	72	97	114
1	52	41	31
2	16	10	5
3	8	4	1
4	4	NA	2
5	NA	1	NA
6	1	NA	NA

Table 33: New counts

	CountGroups	invCategory	NHCategory	NAHRCategory
1	Absence	72	97	114

	CountGroups	invCategory	NHCategory	NAHRCategory
3	Presence	68	51	36
2	Abundance	13	5	3

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

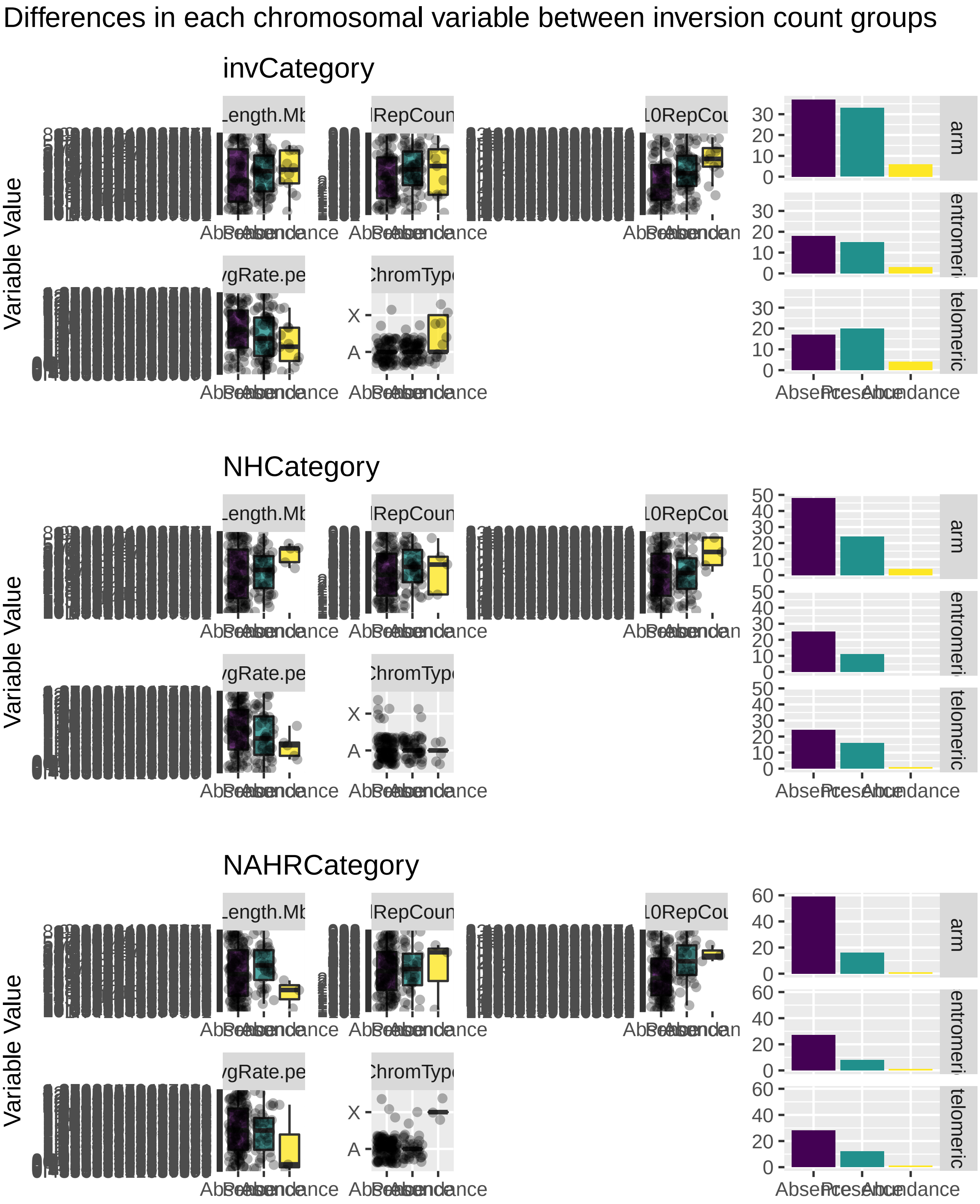


Figure 13: Potential effect of independent variables on the different types of invesions.

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

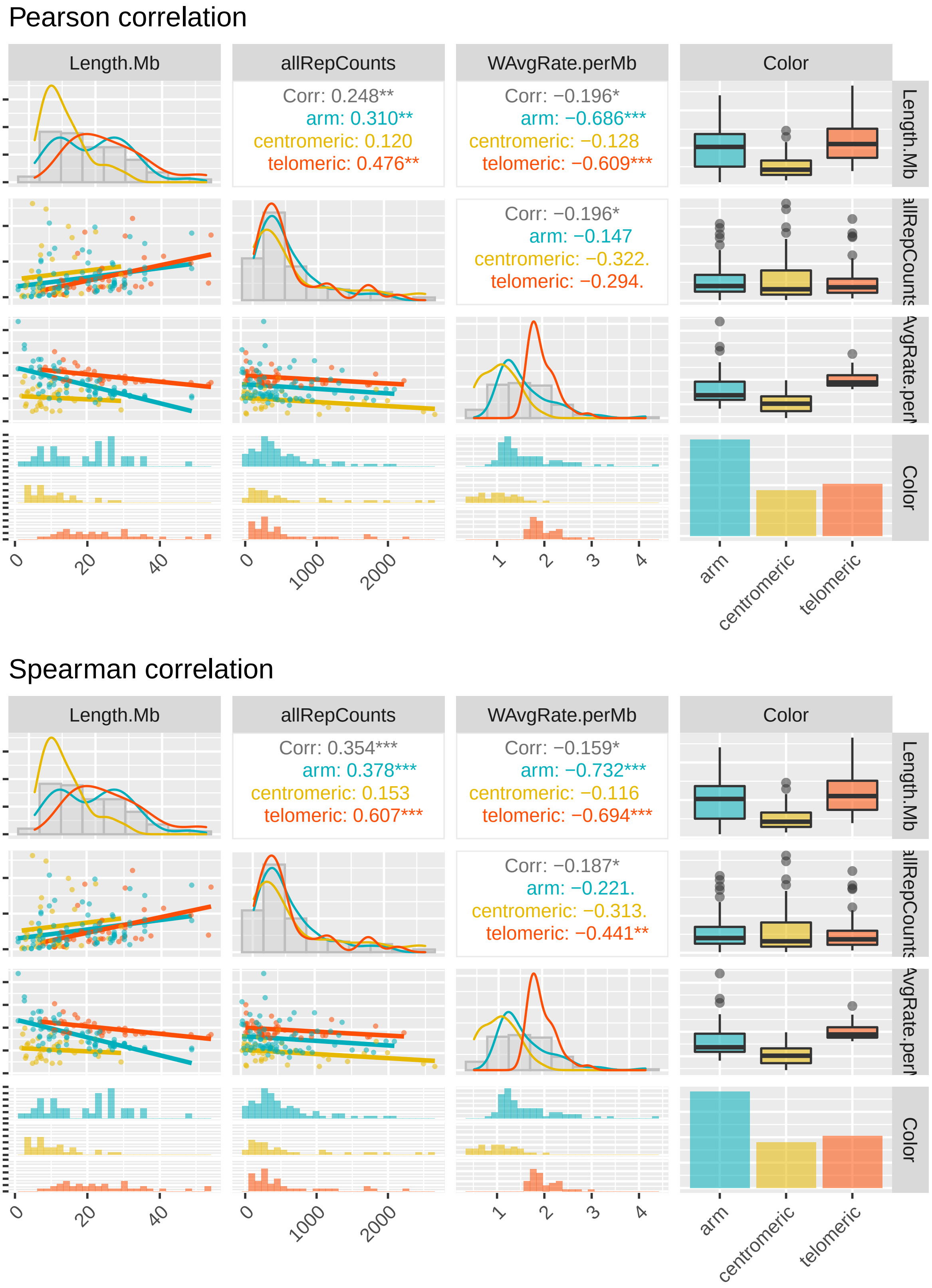


Figure 14: Correlations between variables.

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	1.931714	1	1.389861
allRepCounts	1.105951	1	1.051642
Color	2.573944	2	1.266630
WAvgRate.perMb	2.163202	1	1.470783

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.	1.741944	-1.4979278	16.0000	-0.9932976	0.4356883	-1.8539902
1st Qu.	8.999548	-0.8481517	200.0000	-0.6589042	1.1341848	-0.7230993
Median	16.569122	-0.1704450	374.0000	-0.3426843	1.5359258	-0.0726664
Mean	18.472889	0.0000000	562.5621	0.0000000	1.5808082	0.0000000
3rd Qu.	26.886669	0.7532889	720.0000	0.2861206	1.9196997	0.5486773
Max.	53.232426	3.1120343	2628.0000	3.7536347	4.3762818	4.5259719

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         0.0855532  0.0219058  3.9055
## allRepCounts      0.0005706  0.0003248  1.7568
## Colorcentromeric  0.7419984  0.5657461  1.3115
## Colortelomeric   -0.1002951  0.4654214 -0.2155
## WAvgRate.perMb    0.0819465  0.4279677  0.1915
## ChromTypeX        2.5396945  0.8444356  3.0076
##
## Intercepts:
##                    Value   Std. Error t value
## Absence|Presence    2.0749  1.0269     2.0205
## Presence|Abundance  5.1695  1.1371     4.5464
##
## Residual Deviance: 243.4488
## AIC: 259.4488
```

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

	Value	Std. Error	t value	p value
Length.Mb	0.0855532	0.0219058	3.9055012	0.0000940
allRepCounts	0.0005706	0.0003248	1.7567611	0.0789585
Colorcentromeric	0.7419984	0.5657461	1.3115395	0.1896756
Colortelomeric	-0.1002951	0.4654214	-0.2154932	0.8293828
WAvgRate.perMb	0.0819465	0.4279677	0.1914782	0.8481510
ChromTypeX	2.5396945	0.8444356	3.0075644	0.0026335
Absence|Presence	2.0749097	1.0269418	2.0204745	0.0433342
Presence|Abundance	5.1695153	1.1370579	4.5463958	0.0000055

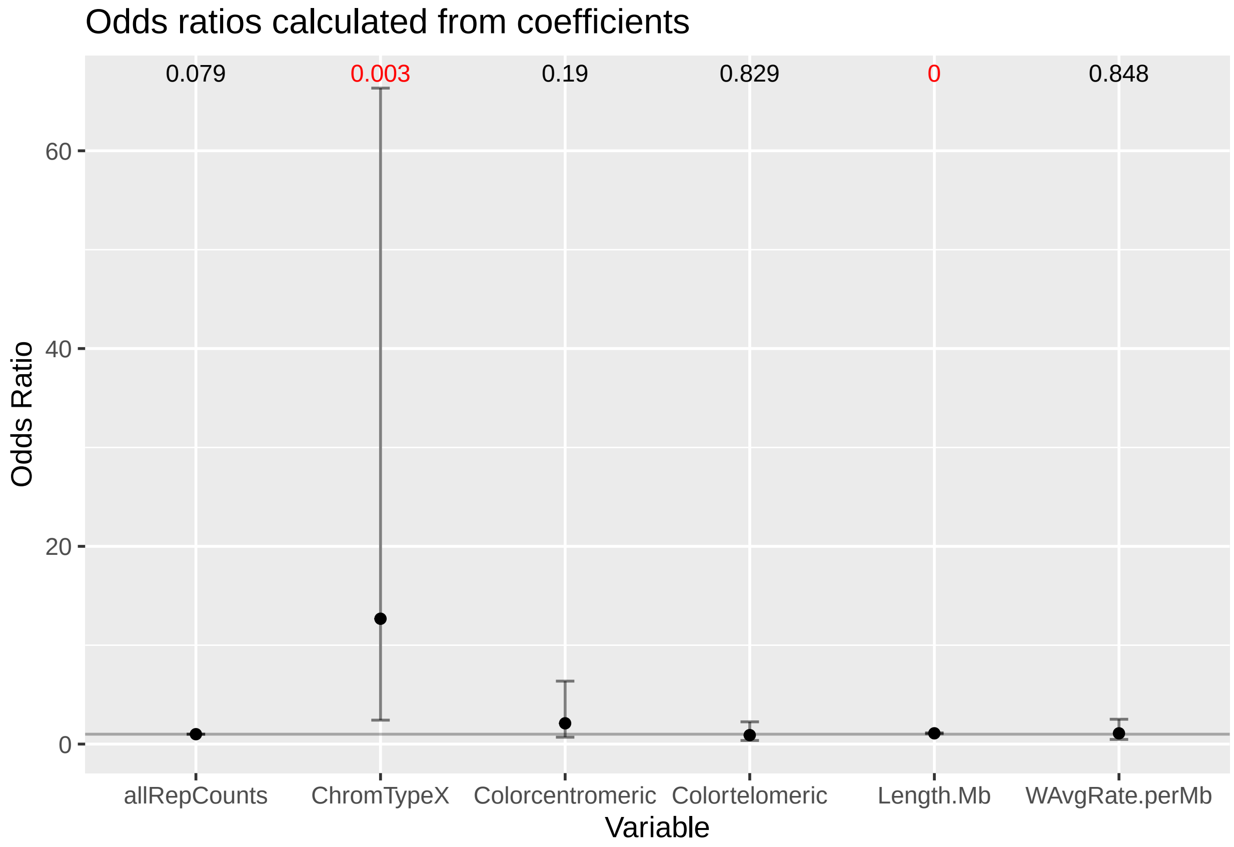
We can also get confidence intervals for the parameter estimates. These can be obtained either by profiling the likelihood function or by using the standard errors and assuming a normal distribution. Note that profiled CIs are not symmetric (although they are usually close to symmetric). If the 95% CI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

	2.5 %	97.5 %
Length.Mb	0.0426186	0.1284877
allRepCounts	-0.0000660	0.0012072
Colorcentromeric	-0.3668436	1.8508404
Colortelomeric	-1.0125043	0.8119141
WAvgRate.perMb	-0.7568548	0.9207478
ChromTypeX	0.8846311	4.1947579

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	1.0893195	1.0435398	1.137107
allRepCounts	1.0005708	0.9999340	1.001208
Colorcentromeric	2.1001281	0.6929180	6.365166
Colortelomeric	0.9045704	0.3633080	2.252215
WAvgRate.perMb	1.0853977	0.4691396	2.511167
ChromTypeX	12.6757977	2.4220907	66.337668

Example of interpretation: "For 1 unit increase in Length.Mb, a window is 1.0893195 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.

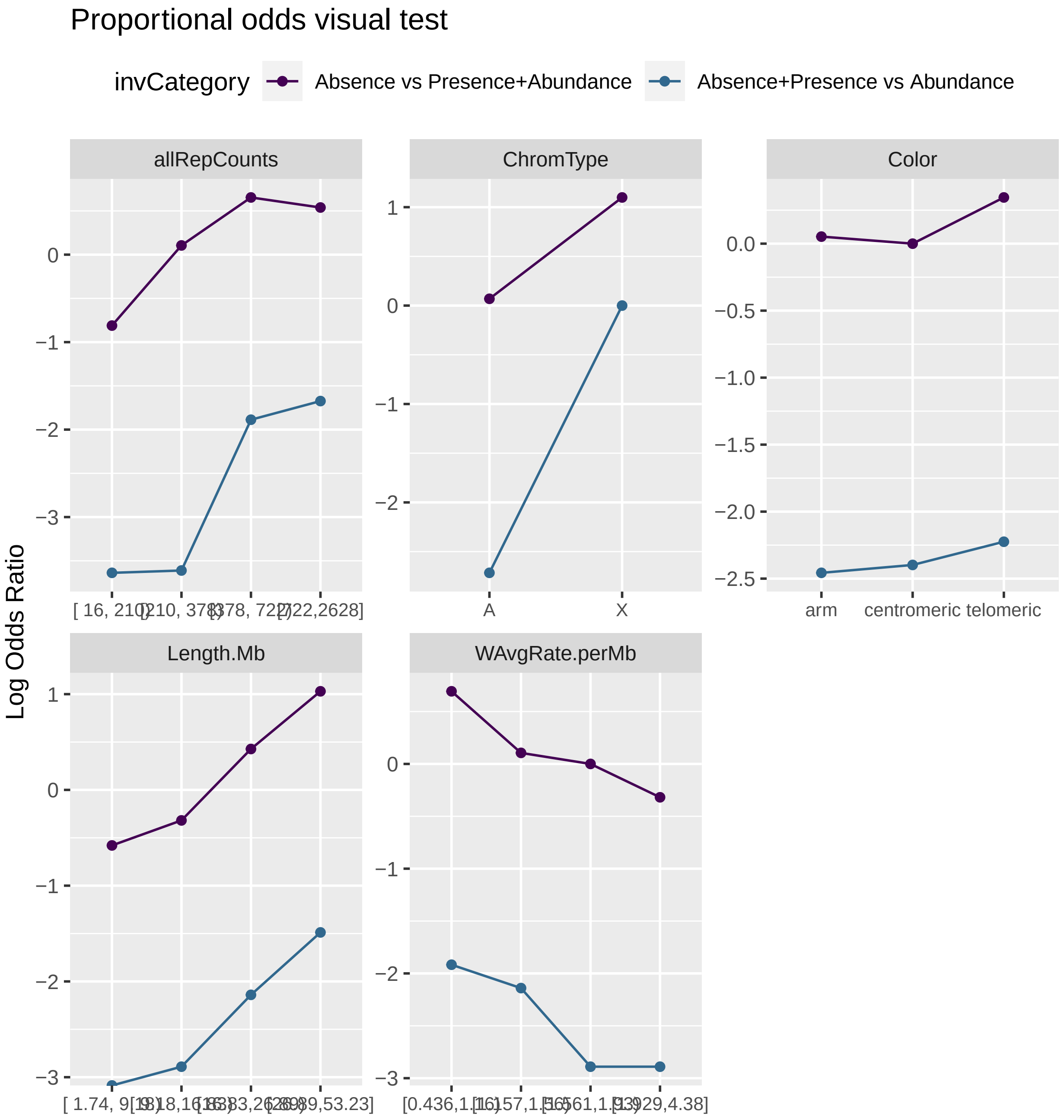
We test the parallel regression assumption with a Brant test:

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

```
## ----------------------------------------------------
## Test for     X2  df  probability
## ----------------------------------------------------
## Omnibus          9.7 6   0.14
## Length.Mb        3.99    1   0.05
## allRepCounts     0.03    1   0.86
## Colorcentromeric 1.46    1   0.23
## Colortelomeric       0.71    1   0.4
## WAvgRate.perMb       0.38    1   0.54
## ChromTypeX       8.46    1   0
## ----------------------------------------------------
##
## H0: Parallel Regression Assumption holds
```

	X2	df	probability
Omnibus	9.6973746	6	0.1379884
Length.Mb	3.9931750	1	0.0456849
allRepCounts	0.0305738	1	0.8611945
Colorcentromeric	1.4616333	1	0.2266704
Colortelomeric	0.7138219	1	0.3981779
WAvgRate.perMb	0.3809101	1	0.5371166
ChromTypeX	8.4597022	1	0.0036310

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.

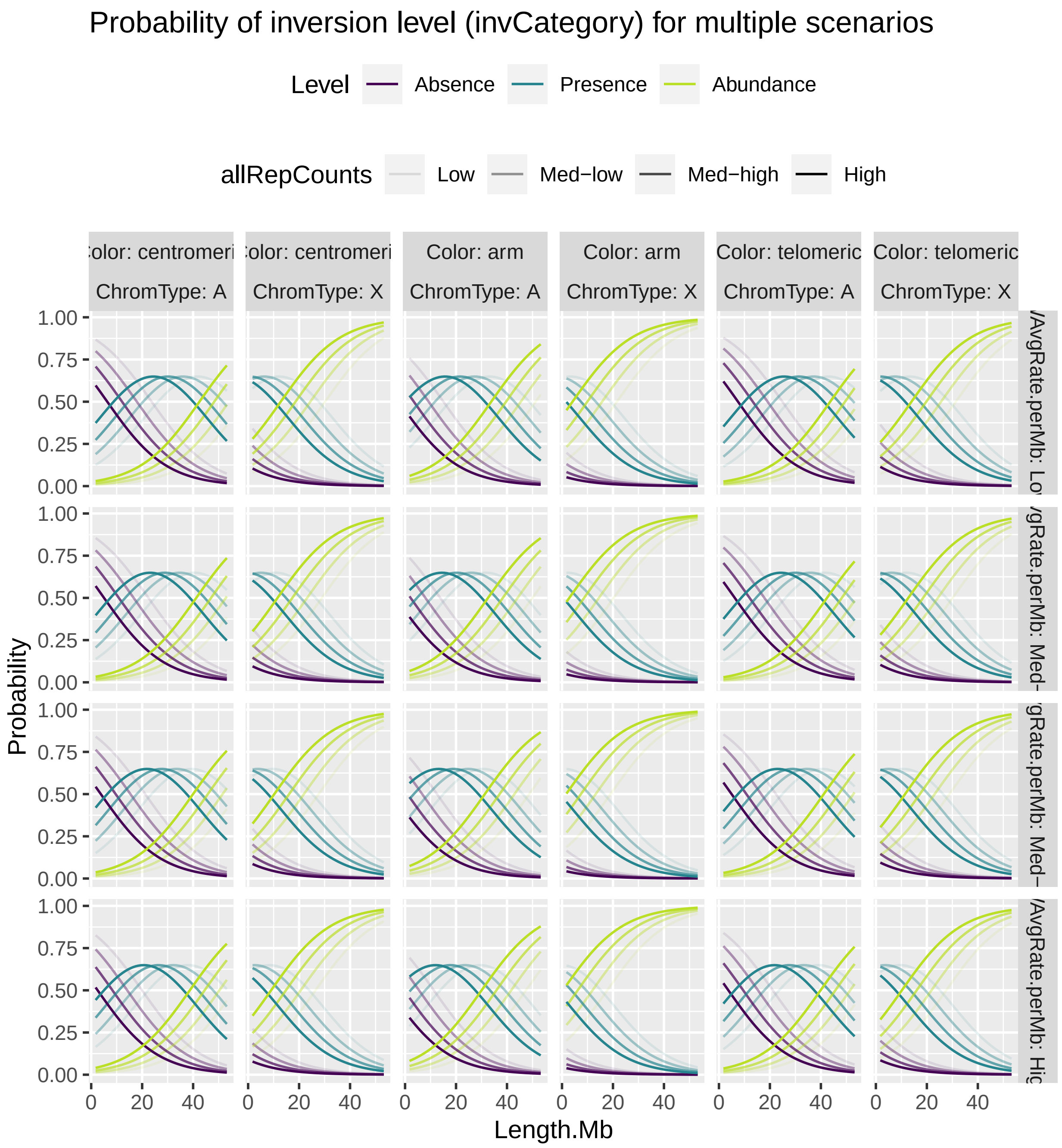


Figure 15: Probabiilty of having 0 to >3 inversions depending on multiple independent variables

NH inversions model

Model fitting

```
## Call:
## polr(formula = myFormula, data = winRegions, Hess = T)
##
## Coefficients:
##                      Value Std. Error  t value
## Length.Mb        0.1075060  0.0243902  4.40775
## allRepCounts    -0.0002631  0.0003707 -0.70980
## Colorcentromeric  0.5714396  0.6080521  0.93979
## Colortelomeric  -0.3476909  0.5270229 -0.65973
## WAvgRate.perMb  -0.0125580  0.5128637 -0.02449
## ChromTypeX      -0.7071296  0.8950270 -0.79007
##
## Intercepts:
##                     Value   Std. Error t value
## Absence|Presence    2.4343  1.1665     2.0868
## Presence|Abundance  5.8100  1.3351     4.3517
##
## Residual Deviance: 200.8628
## AIC: 216.8628
```

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

	Value	Std. Error	t value	p value
Length.Mb	0.1075060	0.0243902	4.4077532	0.0000104
allRepCounts	-0.0002631	0.0003707	-0.7097994	0.4778285
Colorcentromeric	0.5714396	0.6080521	0.9397873	0.3473267
Colortelomeric	-0.3476909	0.5270229	-0.6597264	0.5094294
WAvgRate.perMb	-0.0125580	0.5128637	-0.0244861	0.9804649
ChromTypeX	-0.7071296	0.8950270	-0.7900652	0.4294897
Absence|Presence	2.4343074	1.1665329	2.0867885	0.0369073
Presence|Abundance	5.8100428	1.3351270	4.3516779	0.0000135

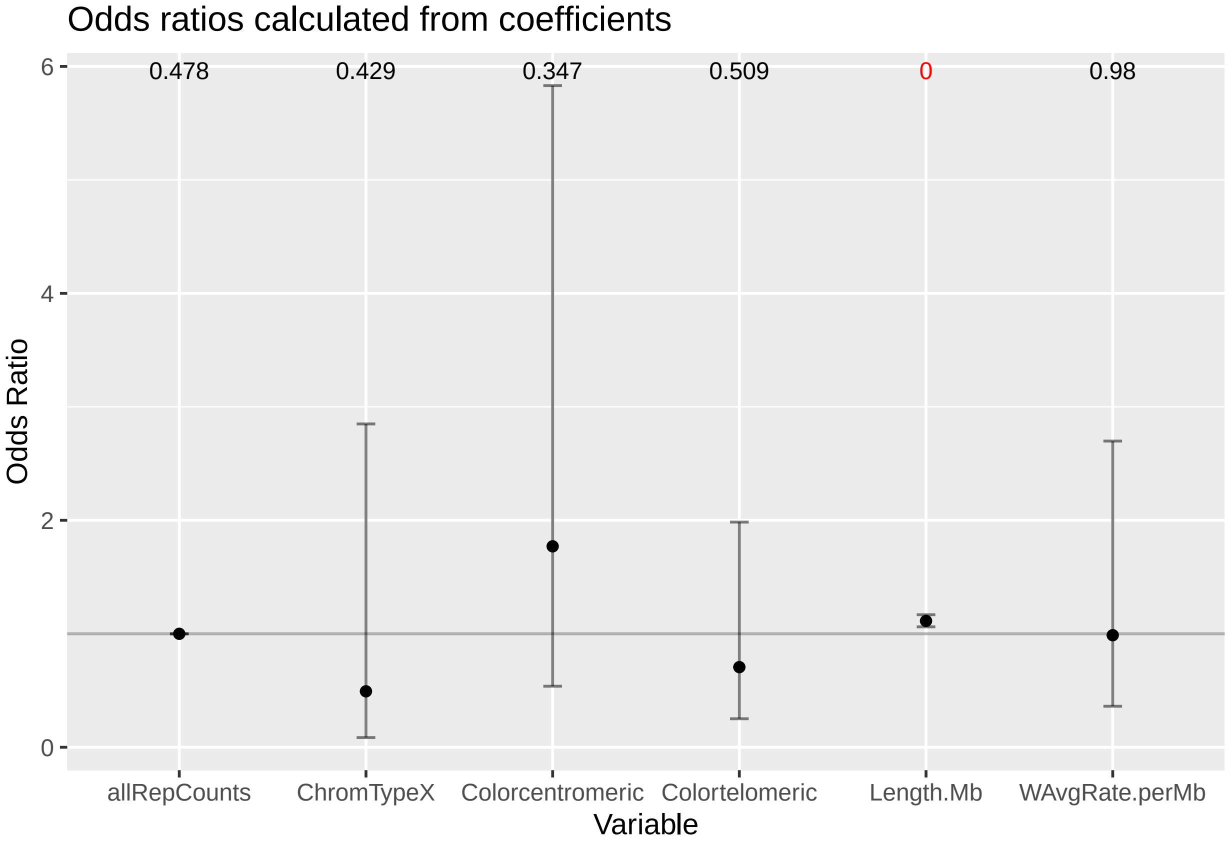
We can also get confidence intervals for the parameter estimates. These can be obtained either by profiling the likelihood function or by using the standard errors and assuming a normal distribution. Note that profiled CIs are not symmetric (although they are usually close to symmetric). If the 95% CI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

	2.5 %	97.5 %
Length.Mb	0.0597021	0.1553099
allRepCounts	-0.0009897	0.0004635
Colorcentromeric	-0.6203206	1.7631998
Colortelomeric	-1.3806368	0.6852550
WAvgRate.perMb	-1.0177524	0.9926363
ChromTypeX	-2.4613502	1.0470910

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	1.1134975	1.0615202	1.168020
allRepCounts	0.9997369	0.9990107	1.000464
Colorcentromeric	1.7708145	0.5377720	5.831066
Colortelomeric	0.7063172	0.2514184	1.984278
WAvgRate.perMb	0.9875205	0.3614063	2.698339
ChromTypeX	0.4930574	0.0853197	2.849350

Example of interpretation: "For 1 unit increase in Length.Mb, a window is 1.1134975 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.

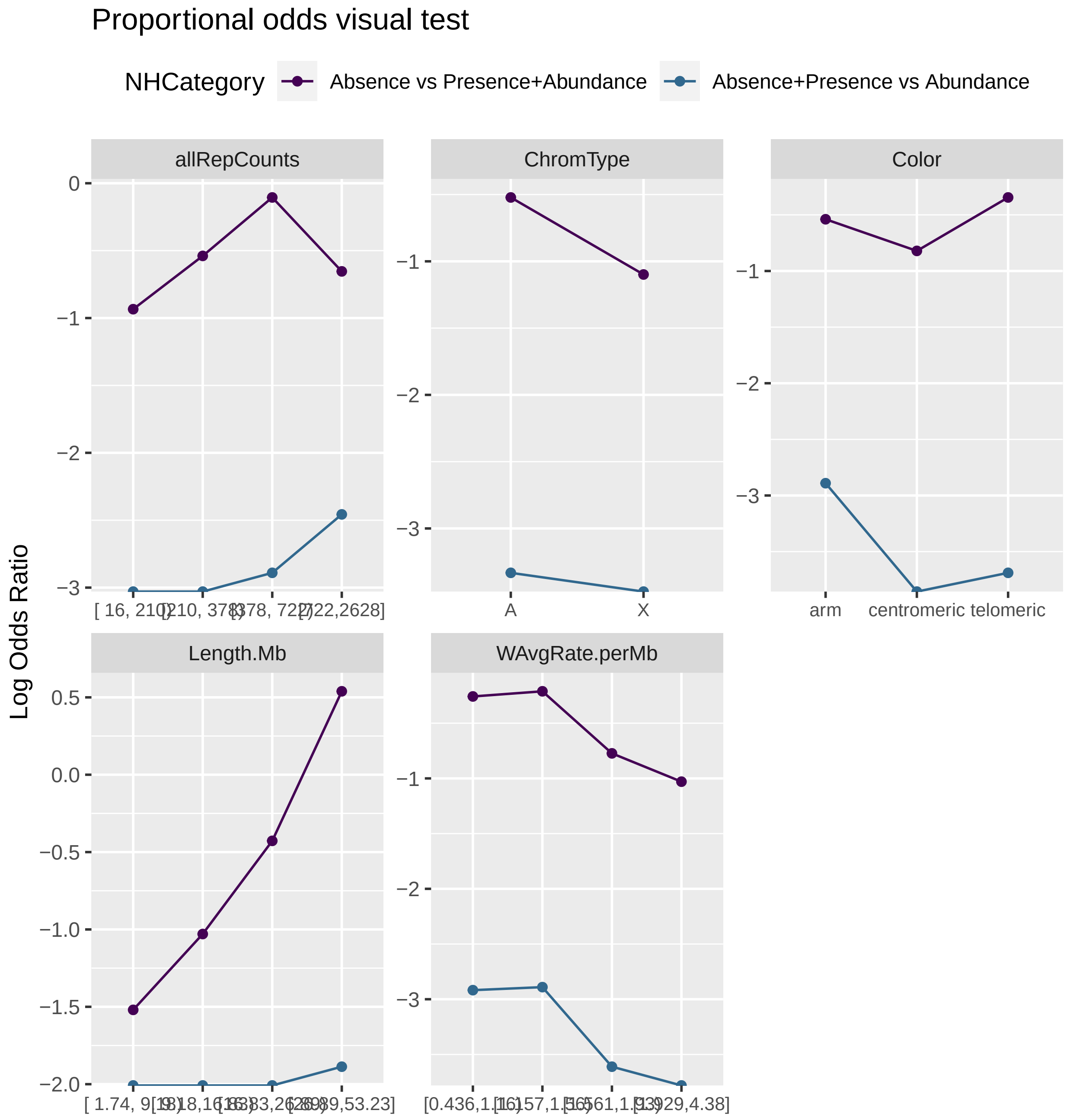
We test the parallel regression assumption with a Brant test:

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

```
## ----------------------------------------------------
## Test for      X2   df   probability
## ----------------------------------------------------
## Omnibus          3.97    6   0.68
## Length.Mb        2.47    1   0.12
## allRepCounts     0.34    1   0.56
## Colorcentromeric 0   1   1
## Colortelomeric       2.53    1   0.11
## WAvgRate.perMb       1.27    1   0.26
## ChromTypeX       0   1   1
## ----------------------------------------------------
##
## H0: Parallel Regression Assumption holds
```

	X2	df	probability
Omnibus	3.9685139	6	0.6809375
Length.Mb	2.4722036	1	0.1158754
allRepCounts	0.3391121	1	0.5603422
Colorcentromeric	0.0000139	1	0.9970302
Colortelomeric	2.5253509	1	0.1120299
WAvgRate.perMb	1.2665008	1	0.2604242
ChromTypeX	0.0000028	1	0.9986654

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.

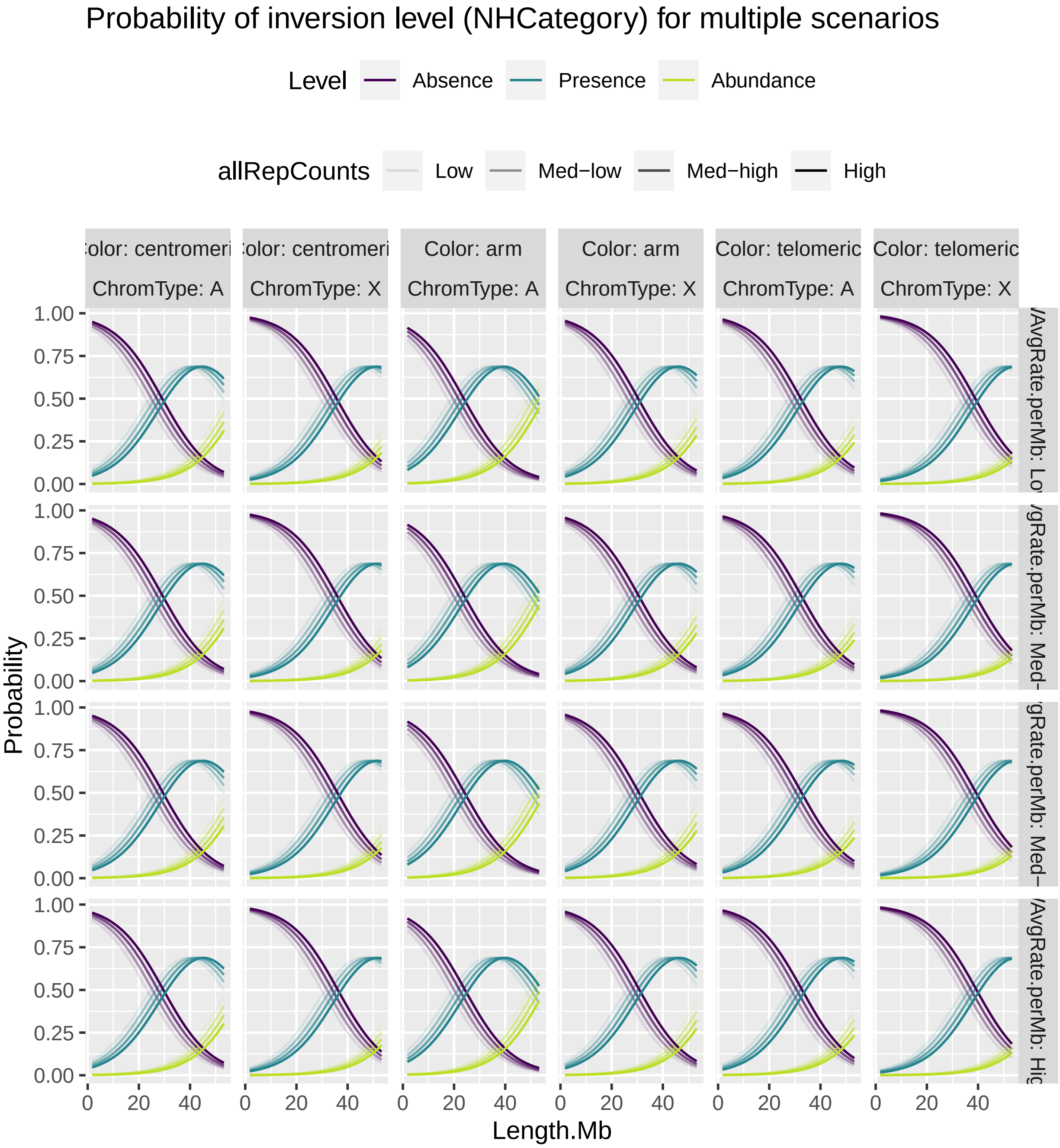


Figure 16: Probabiilty of having 0 to >3 inversions depending on multiple independent variables

NAHR inversions model

Model fitting

```
## Call:
## polr(formula = myFormula, data = winRegions, Hess = T)
##
## Coefficients:
##                       Value Std. Error  t value
## Length.Mb          0.025355  0.0238207  1.06443
## allRepCounts       0.000897  0.0003776  2.37569
## Colorcentromeric   0.279905  0.6904902  0.40537
## Colortelomeric     0.460719  0.5511875  0.83587
## WAvgRate.perMb    -0.023353  0.5433209 -0.04298
## ChromTypeX         3.009461  0.9143551  3.29135
##
## Intercepts:
##                     Value   Std. Error t value
## Absence|Presence    2.4448  1.2532     1.9509
## Presence|Abundance  5.7208  1.4304     3.9995
##
## Residual Deviance: 171.4085
## AIC: 187.4085
```

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

	Value	Std. Error	t value	p value
Length.Mb	0.0253554	0.0238207	1.0644286	0.2871346
allRepCounts	0.0008970	0.0003776	2.3756878	0.0175163
Colorcentromeric	0.2799046	0.6904902	0.4053708	0.6852050
Colortelomeric	0.4607186	0.5511875	0.8358655	0.4032306
WAvgRate.perMb	-0.0233534	0.5433209	-0.0429827	0.9657153
ChromTypeX	3.0094613	0.9143551	3.2913484	0.0009971
Absence|Presence	2.4448136	1.2531525	1.9509307	0.0510653
Presence|Abundance	5.7208263	1.4303878	3.9994932	0.0000635

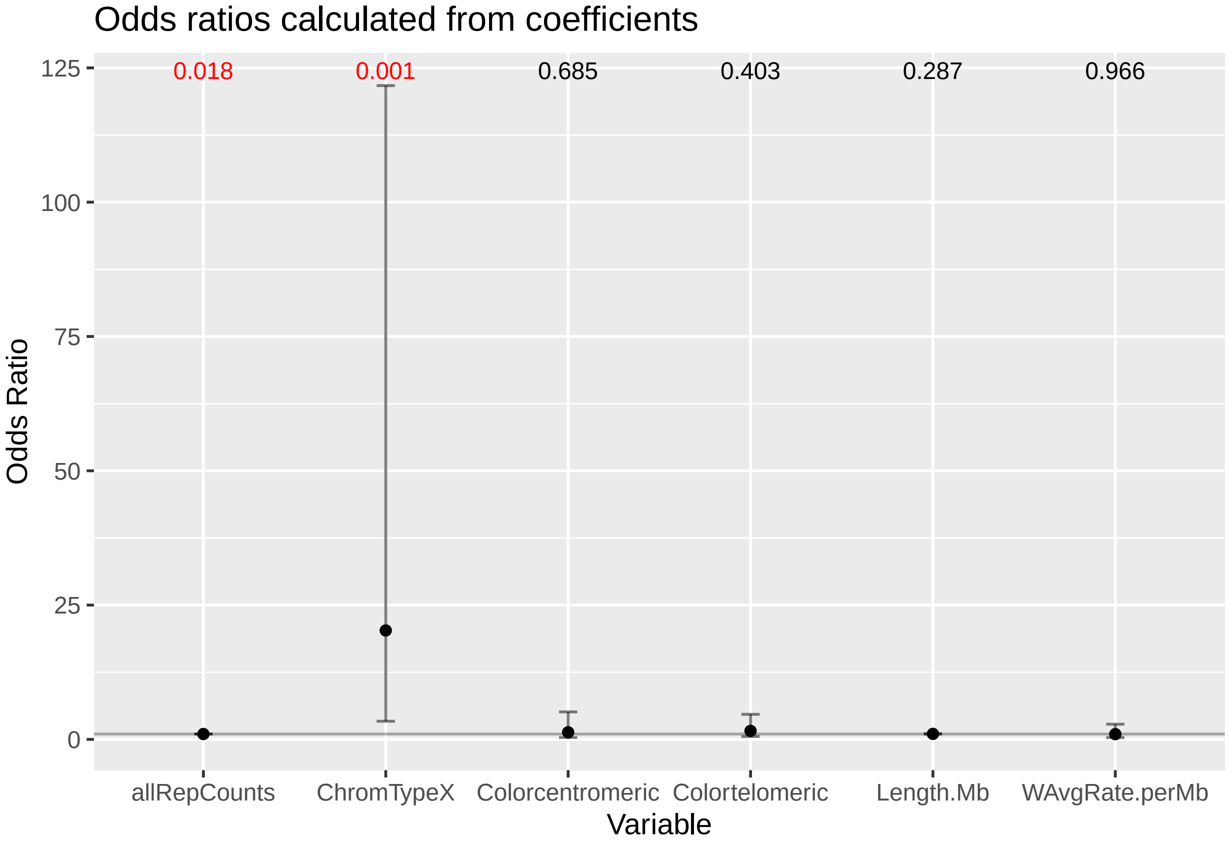
We can also get confidence intervals for the parameter estimates. These can be obtained either by profiling the likelihood function or by using the standard errors and assuming a normal distribution. Note that profiled CIs are not symmetric (although they are usually close to symmetric). If the 95% CI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

	2.5 %	97.5 %
Length.Mb	-0.0213323	0.0720432
allRepCounts	0.0001570	0.0016370
Colorcentromeric	-1.0734314	1.6332406
Colortelomeric	-0.6195890	1.5410262
WAvgRate.perMb	-1.0882429	1.0415361
ChromTypeX	1.2173582	4.8015645

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	1.0256796	0.9788936	1.074702
allRepCounts	1.0008974	1.0001570	1.001638
Colorcentromeric	1.3230036	0.3418335	5.120441
Colortelomeric	1.5852127	0.5381656	4.669380
WAvgRate.perMb	0.9769172	0.3368078	2.833566
ChromTypeX	20.2764747	3.3782512	121.700667

Example of interpretation: "For 1 unit increase in Length.Mb, a window is 1.0256796 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.

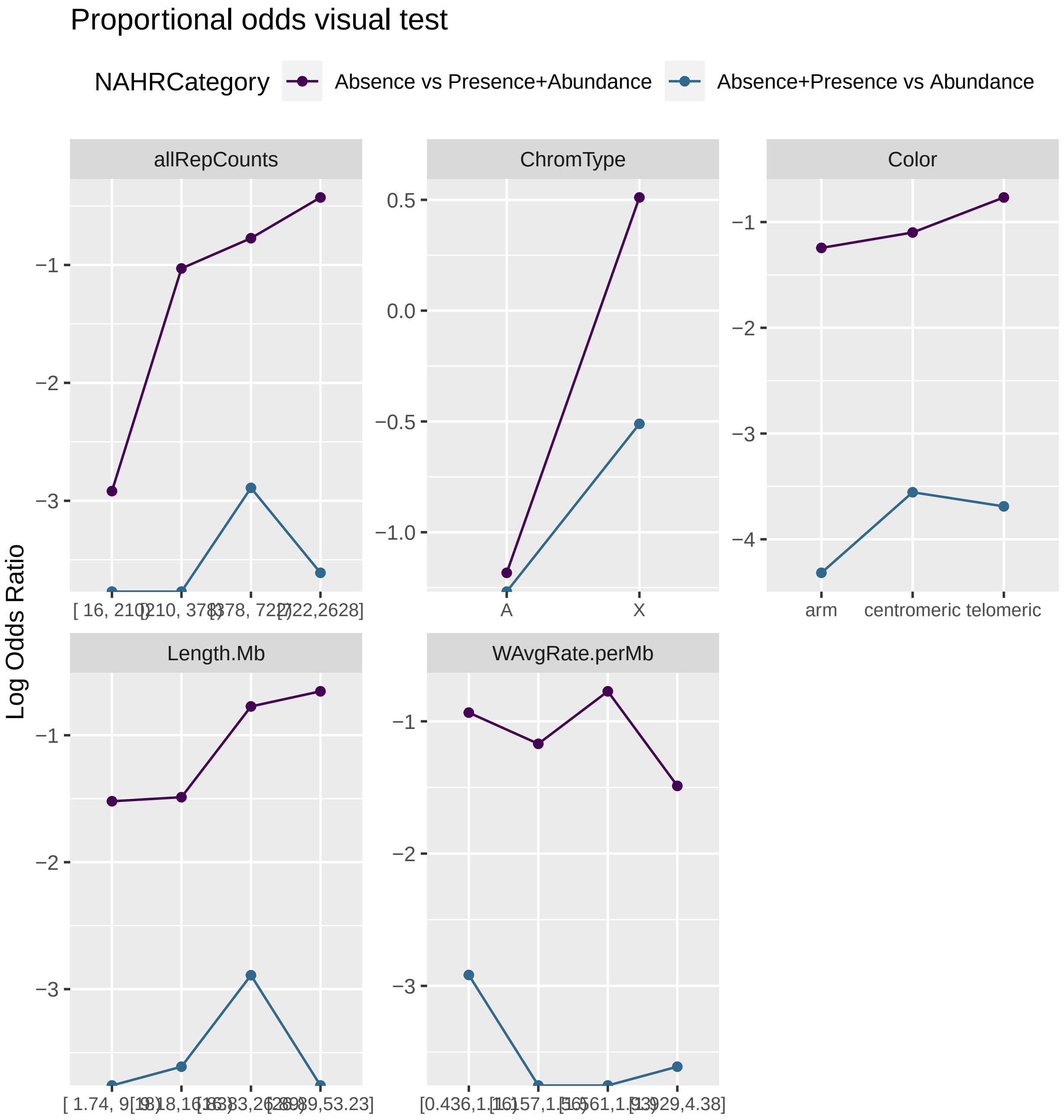
We test the parallel regression assumption with a Brant test:

```
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	4.97e-05	6	1.0000000
Length.Mb	0.00e+00	1	0.9999813
allRepCounts	6.30e-06	1	0.9979996
Colorcentromeric	0.00e+00	1	0.9998494
Colortelomeric	6.00e-07	1	0.9994057
WAvgRate.perMb	2.00e-07	1	0.9996882
ChromTypeX	8.66e-05	1	0.9925736

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.

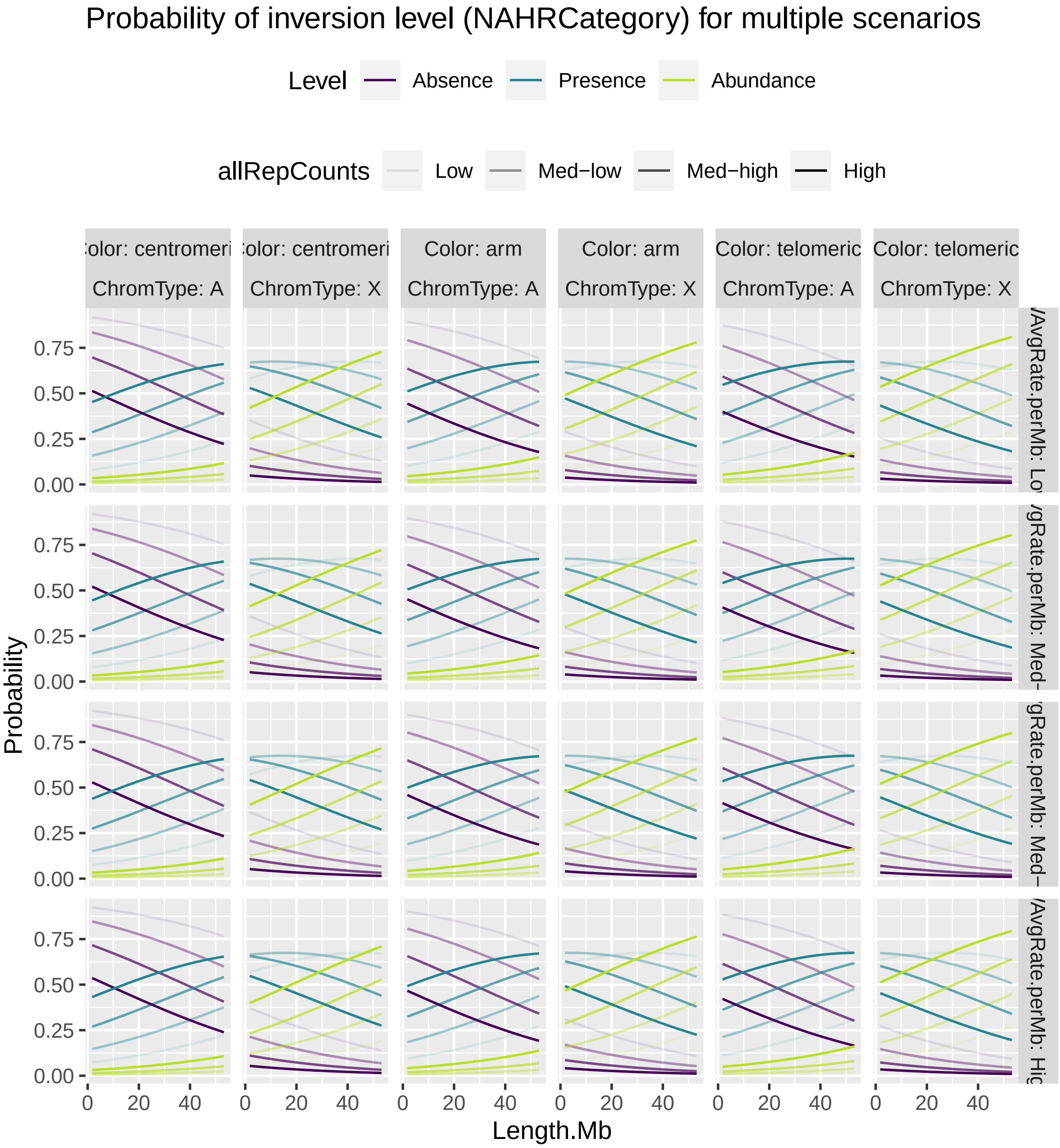


Figure 17: Probabiilty of having 0 to >3 inversions depending on multiple independent variables

Scaled variables

Total inversions model

Model fitting

```
## Call:
## polr(formula = myFormula, data = winRegions, Hess = T)
##
## Coefficients:
##                           Value Std. Error t value
## Length.Mb.Scaled        0.95563     0.2446  3.9068
## allRepCounts.Scaled     0.31400     0.1700  1.8469
## Colorcentromeric        0.74210     0.5657  1.3118
## Colortelomeric         -0.10038     0.4654 -0.2157
## WAvgRate.perMb.Scaled  0.05067     0.2644  0.1917
## ChromTypeX              2.53976     0.8442  3.0085
##
## Intercepts:
##                     Value   Std. Error t value
## Absence|Presence     0.0440  0.2492     0.1764
## Presence|Abundance  3.1386  0.4064     7.7230
##
## Residual Deviance: 243.4488
## AIC: 259.4488
```

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	0.9556254	0.2446064	3.9067881	0.0000935
allRepCounts.Scaled	0.3139957	0.1700102	1.8469231	0.0647583
Colorcentromeric	0.7420964	0.5657004	1.3118188	0.1895813
Colortelomeric	-0.1003783	0.4654245	-0.2156704	0.8292447
WAvgRate.perMb.Scaled	0.0506734	0.2643503	0.1916902	0.8479848
ChromTypeX	2.5397566	0.8441864	3.0085258	0.0026252
Absence|Presence	0.0439542	0.2492129	0.1763720	0.8600017
Presence|Abundance	3.1386350	0.4064019	7.7229833	0.0000000

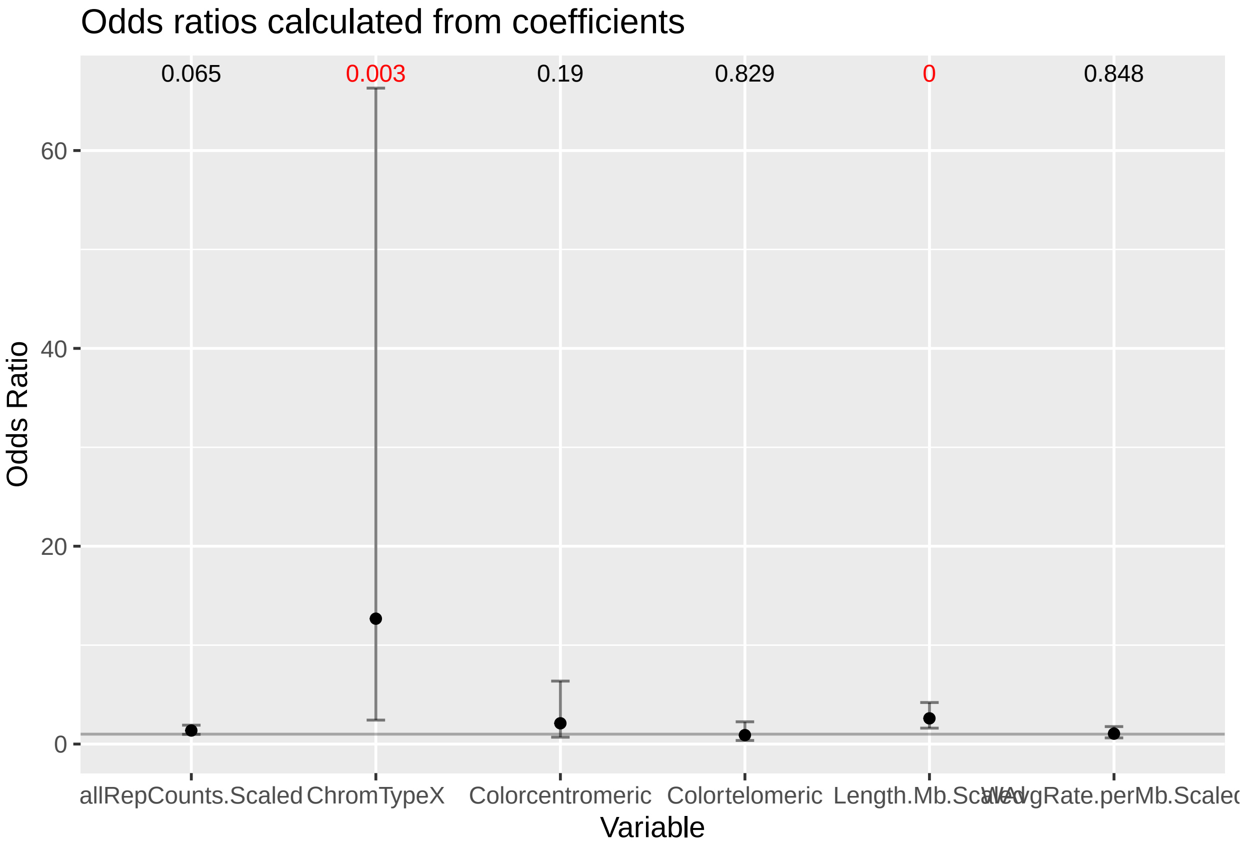
We can also get confidence intervals for the parameter estimates. These can be obtained either by profiling the likelihood function or by using the standard errors and assuming a normal distribution. Note that profiled CIs are not symmetric (although they are usually close to symmetric). If the 95% CI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

	2.5 %	97.5 %
Length.Mb.Scaled	0.4762057	1.4350452
allRepCounts.Scaled	-0.0192181	0.6472095
Colorcentromeric	-0.3666560	1.8508489
Colortelomeric	-1.0125936	0.8118371
WAvgRate.perMb.Scaled	-0.4674437	0.5687904
ChromTypeX	0.8851817	4.1943316

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	2.6002964	1.6099541	4.199835
allRepCounts.Scaled	1.3688838	0.9809654	1.910203
Colorcentromeric	2.1003341	0.6930480	6.365220
Colortelomeric	0.9044952	0.3632756	2.252041
WAvgRate.perMb.Scaled	1.0519792	0.6266020	1.766129
ChromTypeX	12.6765853	2.4234246	66.309394

Example of interpretation: “For 1 unit increase in Length.Mb.Scaled, a window is 2.6002964 times more likely to increase in inversion amount category.”



Proportional odds assessment

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

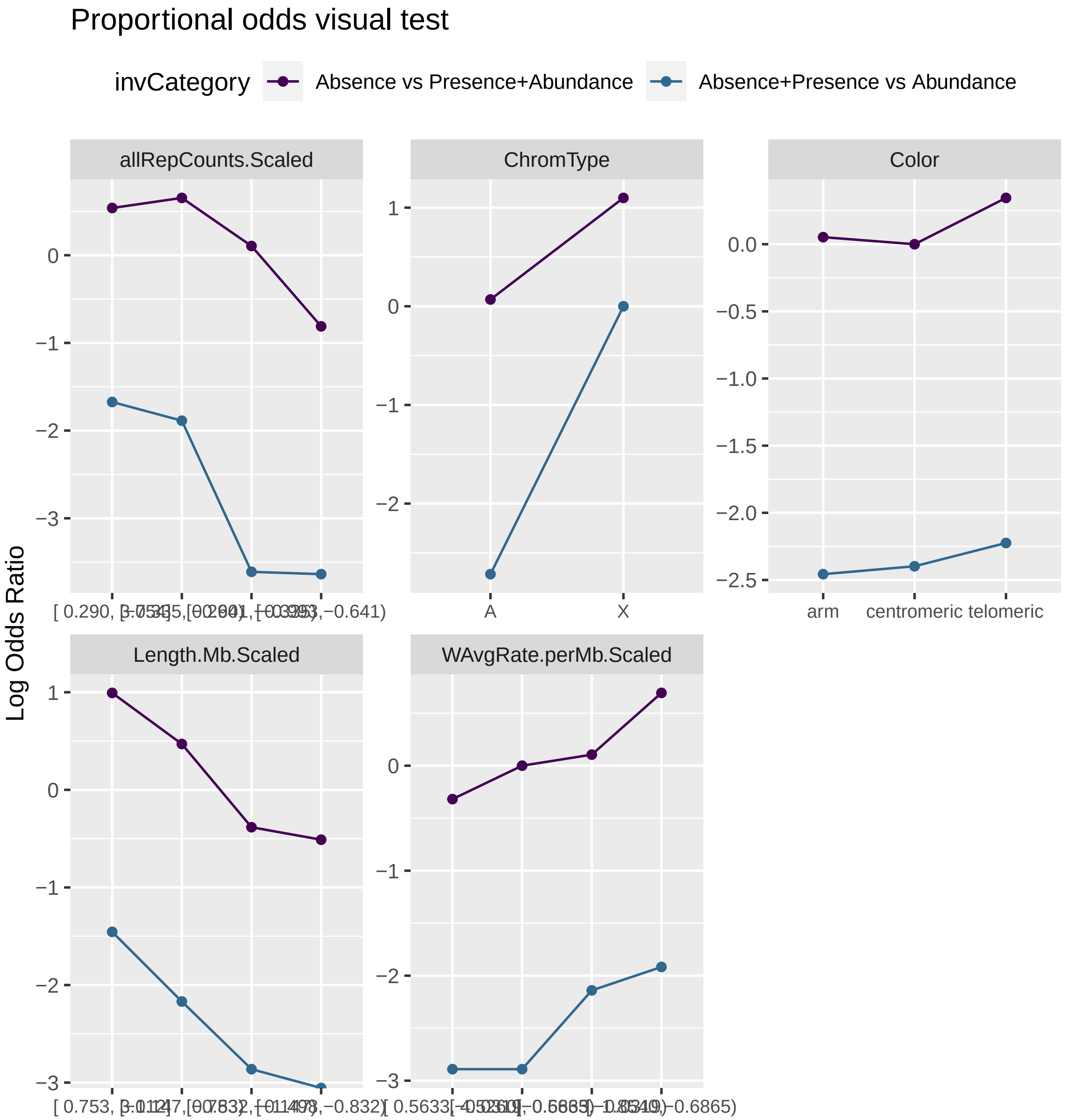
We test the parallel regression assumption with a Brant test:

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

```
## ----------------------------------------------------
## Test for     X2  df  probability
## ----------------------------------------------------
## Omnibus          9.7 6   0.14
## Length.Mb.Scaled 3.99    1   0.05
## allRepCounts.Scaled  0.03    1   0.86
## Colorcentromeric 1.46    1   0.23
## Colortelomeric       0.71    1   0.4
## WAvgRate.perMb.Scaled    0.38    1   0.54
## ChromTypeX       8.46    1   0
## ----------------------------------------------------
##
## H0: Parallel Regression Assumption holds
```

	X2	df	probability
Omnibus	9.6973746	6	0.1379884
Length.Mb.Scaled	3.9931750	1	0.0456849
allRepCounts.Scaled	0.0305738	1	0.8611945
Colorcentromeric	1.4616333	1	0.2266704
Colortelomeric	0.7138219	1	0.3981779
WAvgRate.perMb.Scaled	0.3809101	1	0.5371166
ChromTypeX	8.4597022	1	0.0036310

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.

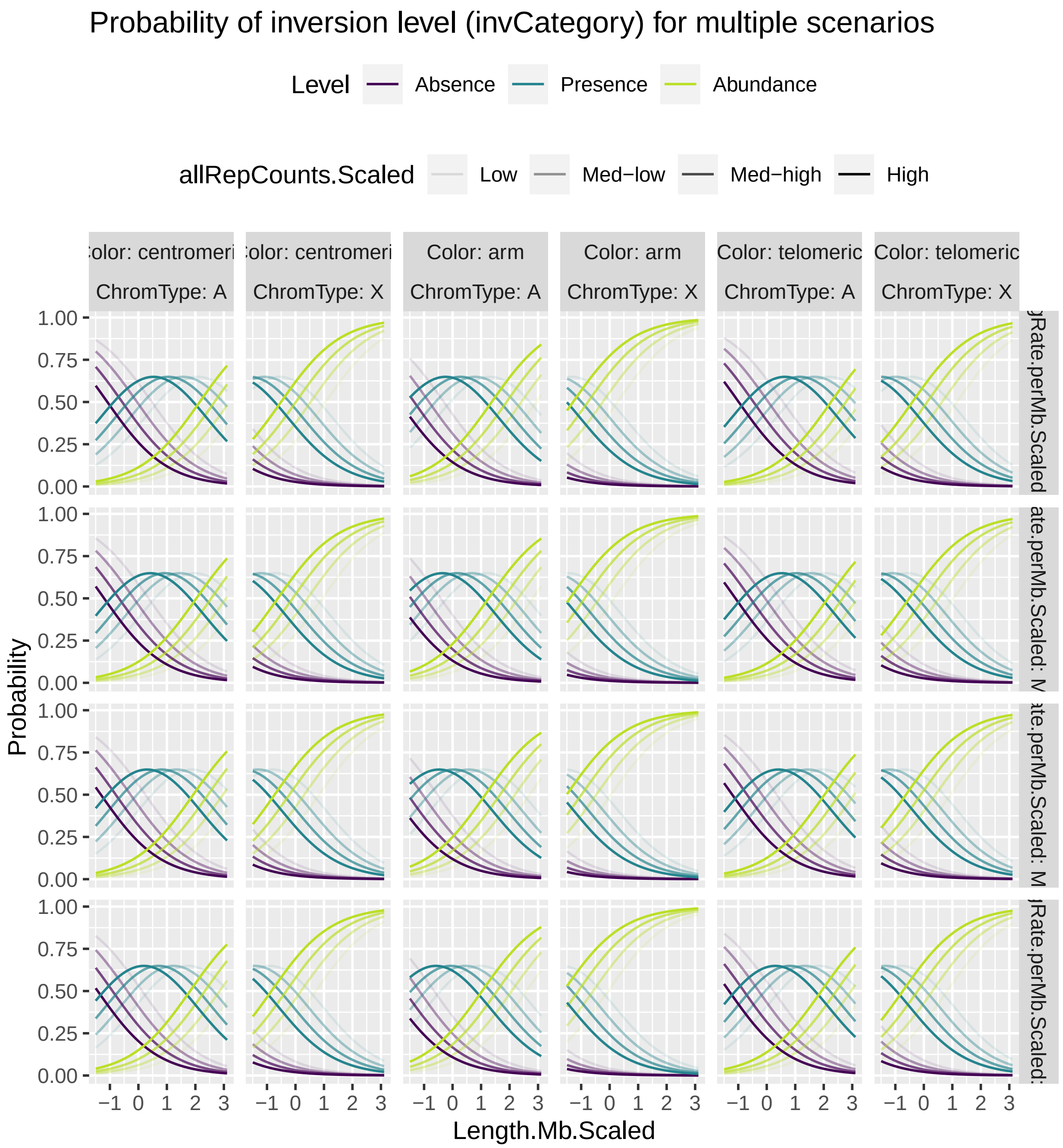


Figure 18: Probabiilty of having 0 to >3 inversions depending on multiple independent variables

NH inversions model

Model fitting

```
## Call:
## polr(formula = myFormula, data = winRegions, Hess = T)
##
## Coefficients:
##                          Value Std. Error  t value
## Length.Mb.Scaled       1.200790     0.2734  4.39127
## allRepCounts.Scaled   -0.144811     0.1950 -0.74255
## Colorcentromeric       0.571442     0.6087  0.93879
## Colortelomeric        -0.347708     0.5270 -0.65975
## WAvgRate.perMb.Scaled -0.007746     0.3167 -0.02446
## ChromTypeX            -0.707115     0.8951 -0.79000
##
## Intercepts:
##                    Value   Std. Error t value
## Absence|Presence    0.6162  0.2721     2.2646
## Presence|Abundance  3.9920  0.5531     7.2174
##
## Residual Deviance: 200.8628
## AIC: 216.8628
```

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	1.2007904	0.2734494	4.3912713	0.0000113
allRepCounts.Scaled	-0.1448108	0.1950193	-0.7425458	0.4577567
Colorcentromeric	0.5714422	0.6087015	0.9387888	0.3478392
Colortelomeric	-0.3477081	0.5270328	-0.6597466	0.5094165
WAvgRate.perMb.Scaled	-0.0077460	0.3167177	-0.0244571	0.9804880
ChromTypeX	-0.7071154	0.8950864	-0.7899969	0.4295296
Absence|Presence	0.6162366	0.2721175	2.2645972	0.0235374
Presence|Abundance	3.9919614	0.5531021	7.2174045	0.0000000

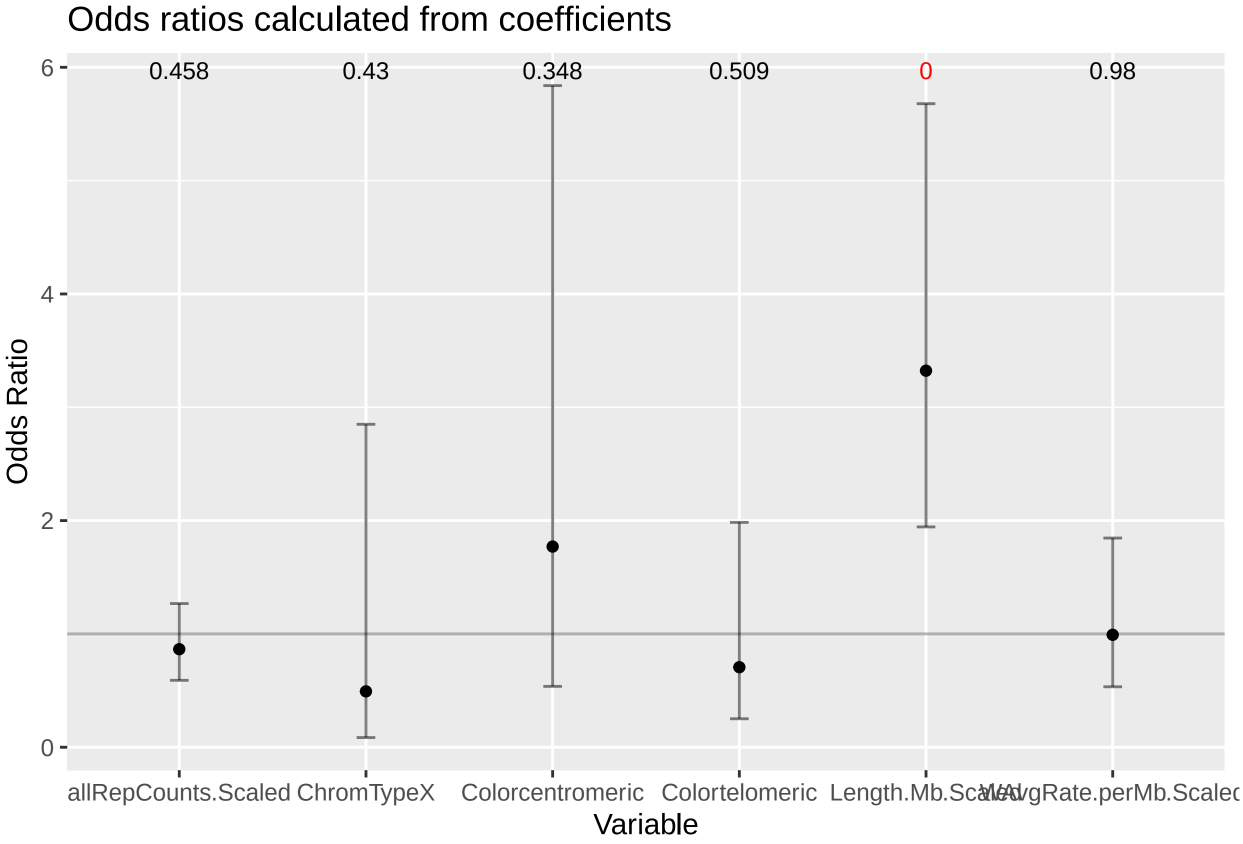
We can also get confidence intervals for the parameter estimates. These can be obtained either by profiling the likelihood function or by using the standard errors and assuming a normal distribution. Note that profiled CIs are not symmetric (although they are usually close to symmetric). If the 95% CI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

	2.5 %	97.5 %
Length.Mb.Scaled	0.6648395	1.7367413
allRepCounts.Scaled	-0.5270417	0.2374201
Colorcentromeric	-0.6215909	1.7644753
Colortelomeric	-1.3806734	0.6852573
WAvgRate.perMb.Scaled	-0.6285013	0.6130093
ChromTypeX	-2.4614524	1.0472216

We convert the coefficients into odds ratios. To get the OR and confidence intervals, we just exponentiate the estimates and confidence intervals (here I used the likelihood confidence intervals).

	Odds Ratio	2.5%	97.5%
Length.Mb.Scaled	3.3227422	1.9441784	5.678808
allRepCounts.Scaled	0.8651860	0.5903488	1.267974
Colorcentromeric	1.7708191	0.5370893	5.838508
Colortelomeric	0.7063050	0.2514092	1.984282
WAvgRate.perMb.Scaled	0.9922839	0.5333906	1.845978
ChromTypeX	0.4930644	0.0853110	2.849722

Example of interpretation: "For 1 unit increase in Length.Mb.Scaled, a window is 3.3227422 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.

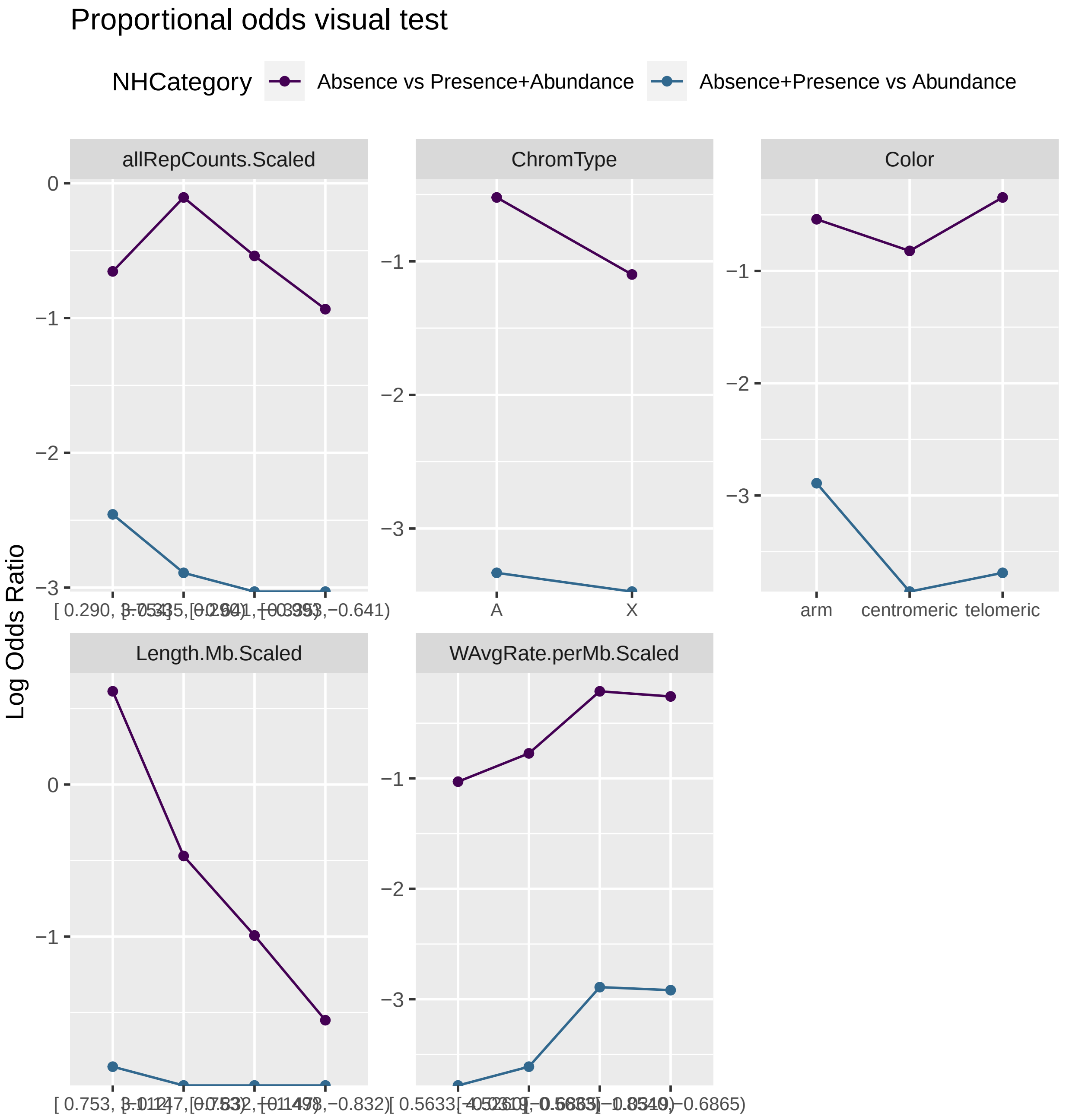
We test the parallel regression assumption with a Brant test:

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

```
## ----------------------------------------------------
## Test for     X2  df  probability
## ----------------------------------------------------
## Omnibus          3.97    6   0.68
## Length.Mb.Scaled 2.47    1   0.12
## allRepCounts.Scaled  0.34    1   0.56
## Colorcentromeric 0   1   1
## Colortelomeric       2.53    1   0.11
## WAvgRate.perMb.Scaled    1.27    1   0.26
## ChromTypeX       0   1   1
## ----------------------------------------------------
##
## H0: Parallel Regression Assumption holds
```

	X2	df	probability
Omnibus	3.9685139	6	0.6809375
Length.Mb.Scaled	2.4722036	1	0.1158754
allRepCounts.Scaled	0.3391121	1	0.5603422
Colorcentromeric	0.0000139	1	0.9970302
Colortelomeric	2.5253509	1	0.1120299
WAvgRate.perMb.Scaled	1.2665008	1	0.2604242
ChromTypeX	0.0000028	1	0.9986654

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.

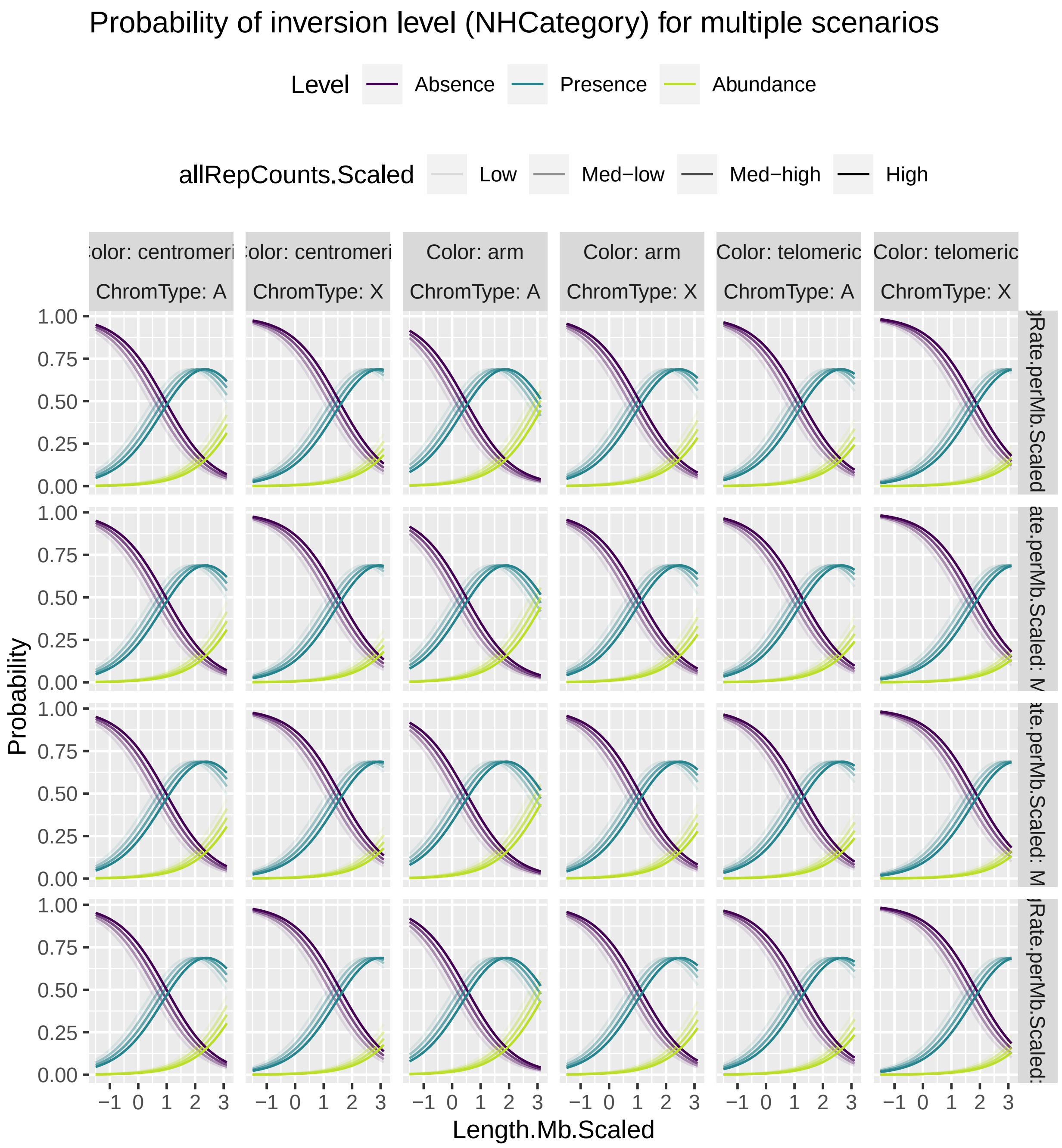


Figure 19: Probabiilty of having 0 to >3 inversions depending on multiple independent variables

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         0.025355  0.0238207  1.06443
## allRepCounts      0.000897  0.0003776  2.37569
## Colorcentromeric  0.279905  0.6904902  0.40537
## Colortelomeric    0.460719  0.5511875  0.83587
## WAvgRate.perMb   -0.023353  0.5433209 -0.04298
## ChromTypeX        3.009461  0.9143551  3.29135
##
## Intercepts:
##                    Value   Std. Error t value
## Absence|Presence    2.4448  1.2532     1.9509
## Presence|Abundance  5.7208  1.4304     3.9995
##
## Residual Deviance: 171.4085
## AIC: 187.4085
```

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

	Value	Std. Error	t value	p value
Length.Mb	0.0253554	0.0238207	1.0644286	0.2871346
allRepCounts	0.0008970	0.0003776	2.3756878	0.0175163
Colorcentromeric	0.2799046	0.6904902	0.4053708	0.6852050
Colortelomeric	0.4607186	0.5511875	0.8358655	0.4032306
WAvgRate.perMb	-0.0233534	0.5433209	-0.0429827	0.9657153
ChromTypeX	3.0094613	0.9143551	3.2913484	0.0009971
Absence|Presence	2.4448136	1.2531525	1.9509307	0.0510653
Presence|Abundance	5.7208263	1.4303878	3.9994932	0.0000635

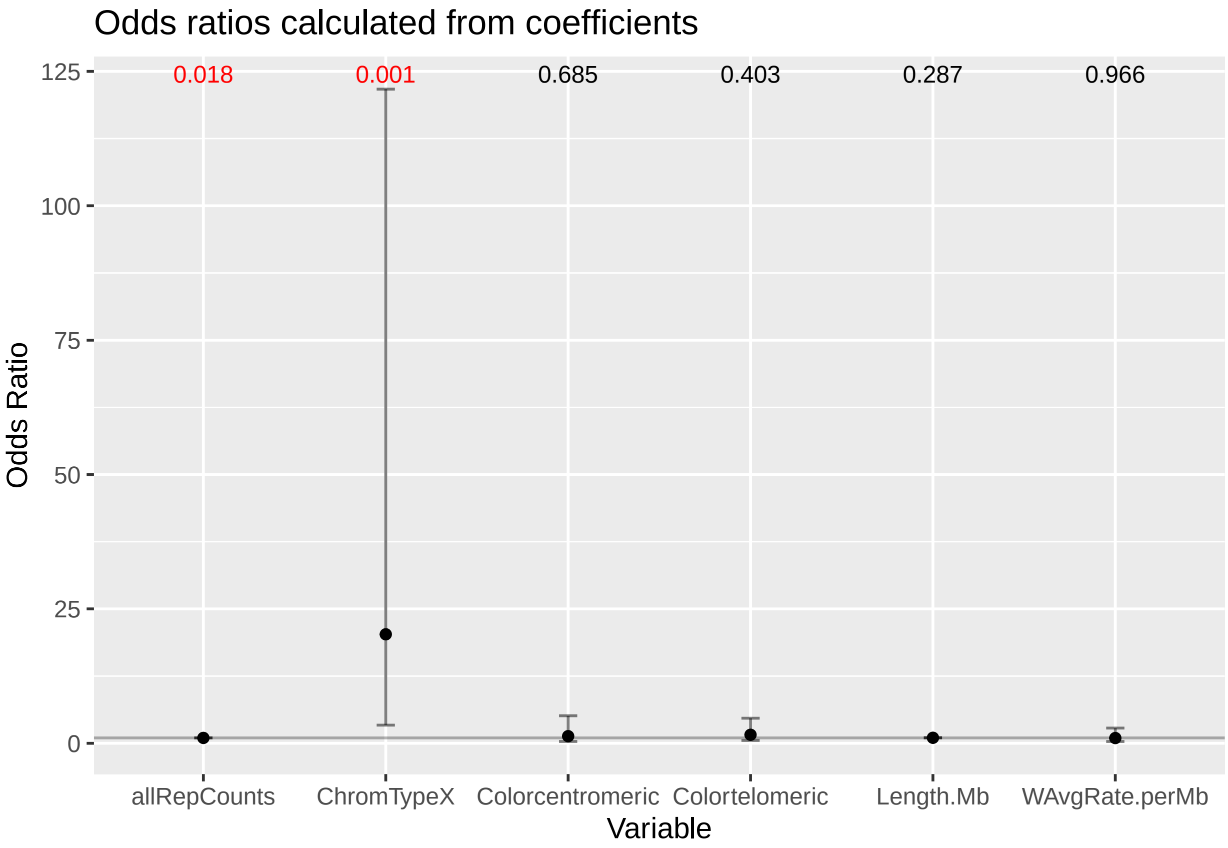
We can also get confidence intervals for the parameter estimates. These can be obtained either by profiling the likelihood function or by using the standard errors and assuming a normal distribution. Note that profiled CIs are not symmetric (although they are usually close to symmetric). If the 95% CI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

	2.5 %	97.5 %
Length.Mb	-0.0213323	0.0720432
allRepCounts	0.0001570	0.0016370
Colorcentromeric	-1.0734314	1.6332406
Colortelomeric	-0.6195890	1.5410262
WAvgRate.perMb	-1.0882429	1.0415361
ChromTypeX	1.2173582	4.8015645

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	1.0256796	0.9788936	1.074702
allRepCounts	1.0008974	1.0001570	1.001638
Colorcentromeric	1.3230036	0.3418335	5.120441
Colortelomeric	1.5852127	0.5381656	4.669380
WAvgRate.perMb	0.9769172	0.3368078	2.833566
ChromTypeX	20.2764747	3.3782512	121.700667

Example of interpretation: "For 1 unit increase in Length.Mb, a window is 1.0256796 times more likely to increase in inversion amount category."



Proportional odds assessment

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

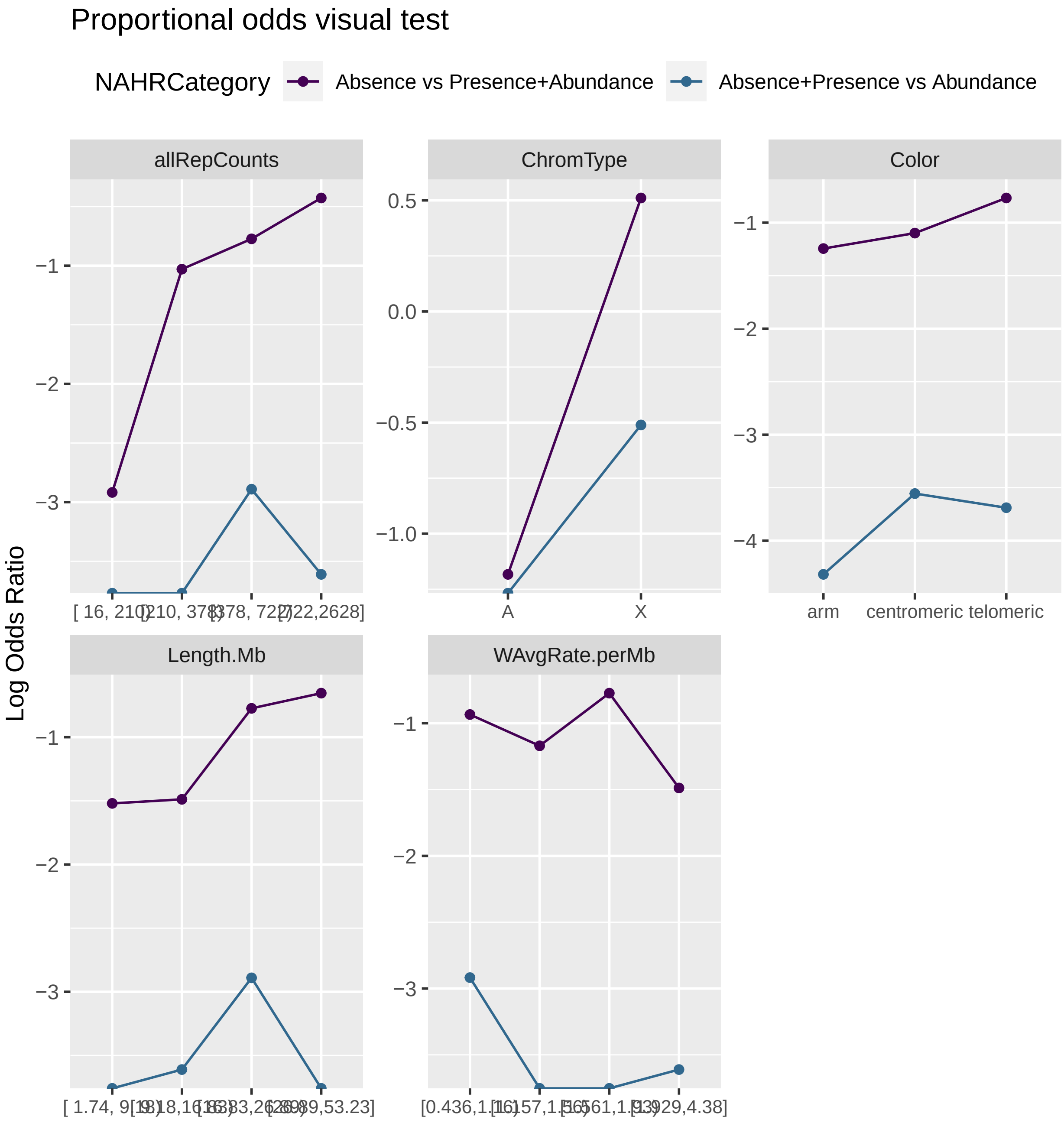
We test the parallel regression assumption with a Brant test:

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

```
## ----------------------------------------------------
## Test for     X2  df  probability
## ----------------------------------------------------
## Omnibus          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	4.97e-05	6	1.0000000
Length.Mb	0.00e+00	1	0.9999813
allRepCounts	6.30e-06	1	0.9979996
Colorcentromeric	0.00e+00	1	0.9998494
Colortelomeric	6.00e-07	1	0.9994057
WAvgRate.perMb	2.00e-07	1	0.9996882
ChromTypeX	8.66e-05	1	0.9925736

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.

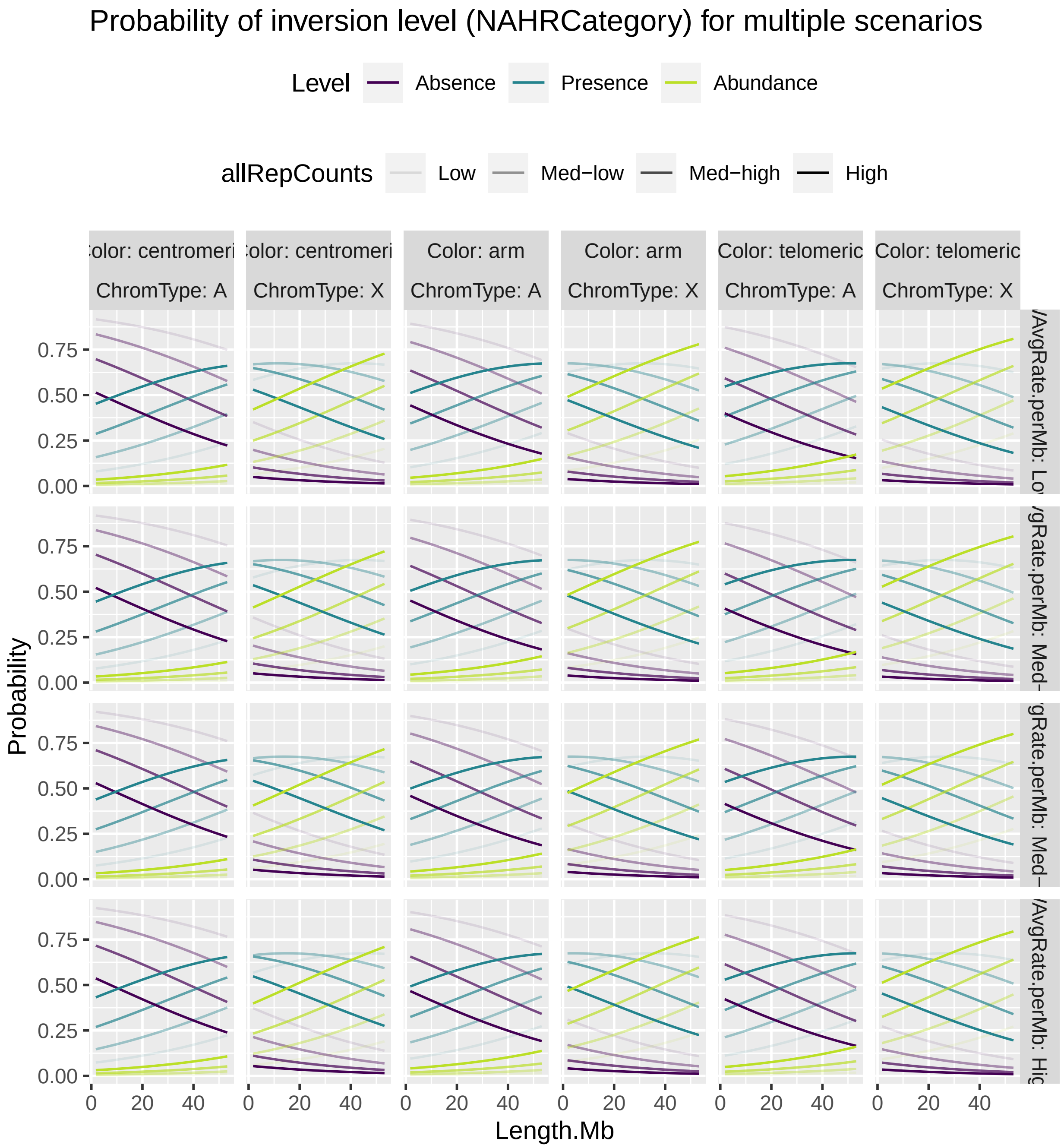


Figure 20: Probabiilty of having 0 to >3 inversions depending on multiple independent variables