Applied Statistical Methods - Solution 5

AUTHOR
Peter von Rohr

PUBLISHED
March 25, 2024

WEBR STATUS
Ready!

Problem 1: Linear Regression on Genomic Information

Use the genomic dataset given at https://charlotte-ngs.github.io/asmasss2024/data/asm_flem_genomic_data.csv to fit a linear regression model of the given observation on the two genomic locations. Assume that both loci have a purely linear effect on the observation.

Tasks

- Read the data from https://charlotte-ngs.github.io/asmasss2024/data/asm_flem_genomic_data.csv

Run Code

```
# read data
s_geno_data <- "https://charlotte-ngs.github.io/asmasss2024/data/asm
df_geno <- read.table(s_geno_data, header = T, sep = ",")
df_geno
```

```
   Animal    SNP.G    SNP.H Observation
1       1 $G_1G_1$ $H_1H_2$         510
2       2 $G_1G_2$ $H_1H_1$         528
3       3 $G_1G_2$ $H_1H_1$         505
4       4 $G_1G_1$ $H_2H_2$         539
5       5 $G_1G_1$ $H_1H_1$         530
6       6 $G_1G_2$ $H_1H_2$         489
7       7 $G_1G_2$ $H_2H_2$         486
8       8 $G_2G_2$ $H_1H_1$         485
9       9 $G_1G_2$ $H_2H_2$         478
10     10 $G_2G_2$ $H_1H_2$         479
11     11 $G_1G_1$ $H_1H_2$         520
12     12 $G_1G_1$ $H_1H_1$         521
13     13 $G_2G_2$ $H_1H_2$         473
14     14 $G_2G_2$ $H_1H_2$         457
15     15 $G_1G_2$ $H_1H_1$         497
16     16 $G_1G_2$ $H_1H_2$         516
17     17 $G_1G_1$ $H_1H_2$         524
18     18 $G_1G_1$ $H_1H_2$         502
19     19 $G_1G_1$ $H_2H_2$         508
20     20 $G_1G_2$ $H_1H_2$         506
```

- Count number of favorable alleles G_1 and H_1

Run Code

```
# counting number of favorable alleles
df_geno$Count.SNP.G <- sapply(df_geno$SNP.G, function(x)
                                               length(grep("1",
                                                 unlist(strsplit(x,
                                 USE.NAMES = F)
df_geno$Count.SNP.H <- sapply(df_geno$SNP.H, function(x)
                                               length(grep("1",
```

```
8                                                    unlist(strsplit(x,
9                                                 USE.NAMES = F)
10  df_geno
```

	Animal	SNP.G	SNP.H	Observation	Count.SNP.G	Count.SNP.H
1	1	G_1G_1	H_1H_2	510	2	1
2	2	G_1G_2	H_1H_1	528	1	2
3	3	G_1G_2	H_1H_1	505	1	2
4	4	G_1G_1	H_2H_2	539	2	0
5	5	G_1G_1	H_1H_1	530	2	2
6	6	G_1G_2	H_1H_2	489	1	1
7	7	G_1G_2	H_2H_2	486	1	0
8	8	G_2G_2	H_1H_1	485	0	2
9	9	G_1G_2	H_2H_2	478	1	0
10	10	G_2G_2	H_1H_2	479	0	1
11	11	G_1G_1	H_1H_2	520	2	1
12	12	G_1G_1	H_1H_1	521	2	2
13	13	G_2G_2	H_1H_2	473	0	1
14	14	G_2G_2	H_1H_2	457	0	1
15	15	G_1G_2	H_1H_1	497	1	2
16	16	G_1G_2	H_1H_2	516	1	1
17	17	G_1G_1	H_1H_2	524	2	1
18	18	G_1G_1	H_1H_2	502	2	1
19	19	G_1G_1	H_2H_2	508	2	0
20	20	G_1G_2	H_1H_2	506	1	1

- Fit regression of observation on count of favorable alleles

Run Code

```
1  # fit linear model of observation on allele counts
2  lm_geno <- lm(Observation ~ Count.SNP.G + Count.SNP.H, data = df_gen
3  summary(lm_geno)
```

```
Call:
lm(formula = Observation ~ Count.SNP.G + Count.SNP.H, data = df_geno)

Residuals:
     Min       1Q   Median       3Q      Max
-18.4643  -8.2468  -0.6883   3.9448  26.9383

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  465.425      7.476  62.252  < 2e-16 ***
Count.SNP.G   23.318      3.861   6.040 1.33e-05 ***
Count.SNP.H    8.403      4.127   2.036   0.0577 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.8 on 17 degrees of freedom
Multiple R-squared:  0.691,    Adjusted R-squared:  0.6546
F-statistic: 19.01 on 2 and 17 DF,  p-value: 4.621e-05
```