Applied Statistical Methods – Solution 1

Peter von Rohr

2020-02-24

Problem 1: Linear Regression

Use the example dataset from the course notes which is used to demonstrate how to fit a regression of the response variable `body weight` (BW) on the predictor variable `breast circumference` (BC). The data is shown in the table below.

Table 1: Dataset for Regression of Body Weight on Breast Circumference for ten Animals

Animal	Breast Circumference	Body Weight
1	176	471
2	177	463
3	178	481
4	179	470
5	179	496
6	180	491
7	181	518
8	182	511
9	183	510
10	184	541

Your Tasks

- Compute the regression coefficient using matrix computations. Use the function `solve()` in R to compute the inverse of a matrix.
- Verify your results using the function `lm` in R.

Solution

- The regression coefficient $\hat{b}_{LS}$ is computed as

$$\hat{b}_{LS} = (X^T X)^{-1} X^T y$$

- The matrix X is

```
##       [,1] [,2]
##  [1,]    1  176
##  [2,]    1  177
##  [3,]    1  178
##  [4,]    1  179
##  [5,]    1  179
##  [6,]    1  180
##  [7,]    1  181
##  [8,]    1  182
##  [9,]    1  183
```

```
## [10,]    1  184
```

- The vector y is

```
##  [1] 471 463 481 470 496 491 518 511 510 541
```

- The regression coefficient is then computed as

```
##  * Intercept:  -1065.115
```

```
##  * Slope:  8.673235
```

- The variance component of the errors is computed as

$$\hat{\sigma^2} = \frac{1}{n-p}\sum_{n=1}^{n} r_i^2$$

```
##  * Error Variance:  122.7997
```

```
##  * Error SD:  11.0815
```

Verifying the results using `lm()`

```
##
## Call:
## lm(formula = `Body Weight` ~ `Breast Circumference`, data = tbl_reg)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.3941  -6.5525  -0.0673   9.3707  13.2594
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            -1065.115    255.483  -4.169 0.003126 **
## `Breast Circumference`     8.673      1.420   6.108 0.000287 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.08 on 8 degrees of freedom
## Multiple R-squared:  0.8234, Adjusted R-squared:  0.8014
## F-statistic: 37.31 on 1 and 8 DF,  p-value: 0.000287
```

Problem 2: Breeding Values

During the lecture the computation of the breeding values for a given genotype was shown for a completely additive locus which means the genotypic value d of the heterozygous genotypes is 0. In this exercise, we want to compute the general solution for the breeding values of all three genotypes under a monogenic model. We are given a single locus G with two alleles G_1 and G_2 which are closely linked to a QTL for a trait of interest. We assume that the population is in Hardy-Weinberg equilibrium at the given locus G. The allele frequencies are

Allele	Frequency
G_1	p
G_2	q

Allele G_1 is the one with a positive effect on the trait of interest. The genotypic values are given in the

following table.

Genotype	Value
G_1G_1	a
G_1G_2	d
G_2G_2	$-a$

Your Task

- Compute the breeding values for all three genotypes G_1G_1, G_1G_2 and G_2G_2.
- Verify the results presented in the lecture by setting $d = 0$ in the breeding values you computed before.

Solution

The breeding value for an animal with a given genotype is defined as two times the deviation of a large number of progeny from the population mean. Based on that definition, we first compute the population mean

$$\begin{aligned}\mu &= f(G_1G_1) * a + f(G_1G_2) * d + f(G_2G_2)(-a) \\ &= p^2 * a + 2pq * d - q^2 * a \\ &= (p^2 - q^2) * a + 2pqd \\ &= (p - q)a + 2pqd \end{aligned} \tag{1}$$

For each of the genotypes G_1G_1, G_1G_2 and G_2G_2 we compute the expected genotypic value of the offspring. Taking the difference from the expected genotypic value of the offspring of animals with the different genotypes and multiply that differerence with two yields the breeding value.

Genotype G_1G_1: The following table gives an overview over the genotype frequencies of the offspring of a parent with a G_1G_1 genotype

	Sire	
	G_1	G_2
Dam		
G_1	$f(G_1G_1) = p$	$f(G_1G_2) = q$

The expected genotypic value μ_{11} of the offspring of G_1G_1

$$\mu_{11} = p * a + q * d \tag{2}$$

The breeding value BV_{11} of an animal with genotype G_1G_1

$$\begin{aligned} BV_{11} &= 2 * (\mu_{11} - \mu) \\ &= 2 * (pa + qd - [(p - q)a + 2pqd]) \\ &= 2 * (pa + qd - pa + qa - 2pqd) \\ &= 2q * (a + (1 - 2p)d) \\ &= 2q * (a + (q - p)d) \\ &= 2q\alpha \end{aligned} \tag{3}$$

Genotype G_1G_2: The table with the offspring genotype frequencies

	Sire	
	G_1	G_2
Dam		
G_1	$f(G_1G_1) = 0.5p$	$f(G_1G_2) = 0.5q$
G_2	$f(G_2G_1) = 0.5p$	$f(G_2G_2) = 0.5q$

The expected genotypic value μ_{12} of the offpring of a G_1G_2 parent is

$$\mu_{12} = 0.5p * a + 0.5(p+q) * d + 0.5q * (-a) = 0.5pa + 0.5d - 0.5qa \tag{4}$$

The breeding value BV_{12} is

$$\begin{aligned} BV_{12} &= 2 * (\mu_{12} - \mu) \\ &= 2 * (0.5pa + 0.5d - 0.5qa - [(p-q)a + 2pqd]) \\ &= 2 * (0.5qa - 0.5pa + 0.5d - 2pqd) \\ &= (q-p)a + (1-4pq)d \\ &= (q-p)a + (p^2 + q^2 + 2pq - 4pq)d \\ &= (q-p)a + (p-q)^2 d \\ &= (q-p)(a + (q-p)d) \\ &= (q-p)\alpha \end{aligned} \tag{5}$$

Genotype G_2G_2: The table with the offspring genotype frequencies

	Sire	
	G_1	G_2
Dam		
G_2	$f(G_2G_1) = p$	$f(G_2G_2) = q$

The expected genotypic value μ_{22} of the offpring of a G_2G_2 parent is

$$\mu_{22} = p * d + q * (-a) = pd - qa \tag{6}$$

The breeding value BV_{12} is

$$\begin{aligned} BV_{22} &= 2 * (\mu_{22} - \mu) \\ &= 2 * (pd - qa - [(p-q)a + 2pqd]) \\ &= 2 * (pd - pa - 2pqd) \\ &= 2 * (-pa + p(1-2q)d) \\ &= -2p * (a + (q-p)d) \\ &= -2p\alpha \end{aligned} \tag{7}$$

In summary the breeding values are

Genotype	Breeding Value

G_1G_1	$2q\alpha$
G_1G_2	$(q-p)\alpha$
G_2G_2	$-2p\alpha$

All breeding values depend on $\alpha = a + (q-p)d$. For purely additive loci, $d = 0$ and therefore $\alpha = a$. Then the breeding values simplify to

Genotype	Breeding Value
G_1G_1	$2qa$
G_1G_2	$(q-p)a$
G_2G_2	$-2pa$

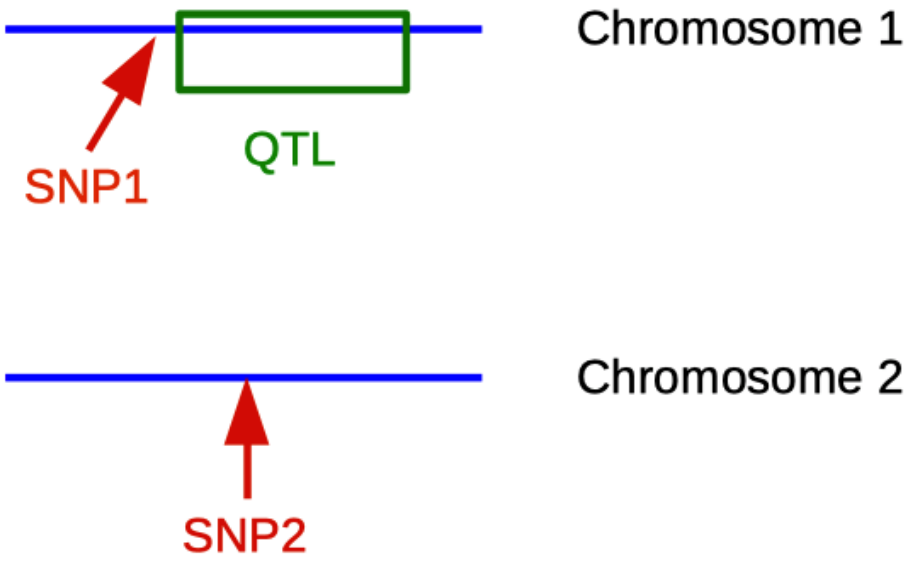

Figure 1: Linkage Between an SNP and a QTL and an independent SNP on a different Chromosome

Problem 3: Linkage Between SNP and QTL

In a population of breeding animals, we are given a trait of interest which is determined by a QTL Q on chromosome 1. QTL Q is modelled as a bi-allelic QTL with alleles Q_1 and Q_2. Furthermore we have genotyped our population for two SNPs R and S with two alleles each. One of the SNPs is on chromosome 1 and is closely linked to Q. The other SNP is on chromosome 2 and is unlinked. Figure 1 shows the situation in a diagram.

Based on the following small dataset, determine which of the two SNPs R and/or S is linked to QTL Q.

From the above table it might be difficult to decide which SNP is linked to the QTL. Plotting the data may help. Showing the observations as a function of the genotypes leads to Figure 2.

Your Tasks

- Determine which of the two SNPs R or S is closely linked to the QTL
- Estimate a value for a obtained based on the data
- Try to fit a linear model through the genotypes that SNP which is linked to the QTL using the `lm()` function. The genotype data is available from

`https://charlotte-ngs.github.io/GELASMSS2020/ex/w02/asm_w02_ex01_p02_genodatafile.csv`

Table 6: Dataset showing linkage between SNP and QTL

SNP R	SNP S	Observation
R_2R_2	S_1S_1	23.17
R_2R_2	S_2S_2	-27.04
R_1R_2	S_1S_2	-2.79
R_1R_2	S_2S_2	-19.54
R_1R_2	S_2S_2	-24.05
R_1R_2	S_1S_1	25.84
R_1R_2	S_1S_2	-0.36
R_1R_1	S_2S_2	-23.34
R_2R_2	S_1S_2	1.38
R_1R_1	S_1S_2	-1.60
R_1R_2	S_1S_2	-2.97
R_2R_2	S_1S_2	-1.39

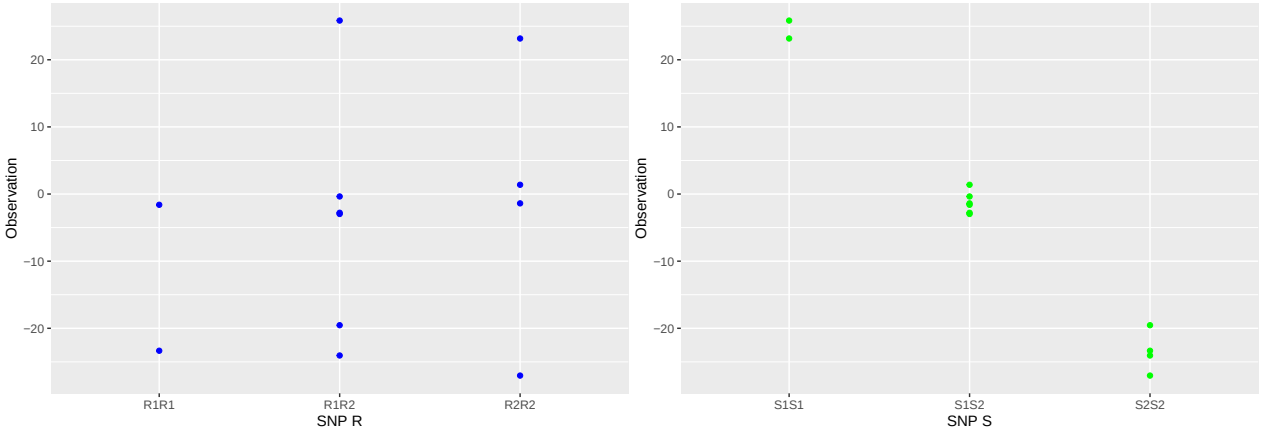

Figure 2: Observations Grouped by SNP Genotypes

Solution

1. Based on the plot shown above, the SNP S is linked to the QTL.
2. Fit the linear model of the observations

```
s_asm_w02_ex01_p02_genodatafile <-
  "https://charlotte-ngs.github.io/GELASMSS2019/ex/w02/asm_w02_ex01_p02_genodatafile.csv"
tbl_all_data_ascii <- readr::read_csv(file = s_asm_w02_ex01_p02_genodatafile)
```

```
## Parsed with column specification:
## cols(
##   `SNP R` = col_character(),
##   `SNP S` = col_character(),
##   Observation = col_double()
## )
```

```
tbl_all_data_ascii$`SNP R` <- as.factor(tbl_all_data_ascii$`SNP R`)
tbl_all_data_ascii$`SNP S` <- as.factor(tbl_all_data_ascii$`SNP S`)
lm_fit_geno_snp_r <- lm(Observation ~ 0 + `SNP R`, data = tbl_all_data_ascii)
summary(lm_fit_geno_snp_r)
```

```
##
## Call:
## lm(formula = Observation ~ 0 + `SNP R`, data = tbl_all_data_ascii)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -25.1517  -7.0258  -0.9758   6.1017  25.0783
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## `SNP R`R1R1   2.2300     9.5965   0.232    0.821
## `SNP R`R1R2   0.8817     6.7858   0.130    0.899
## `SNP R`R2R2 -16.7767     9.5965  -1.748    0.114
##
## Residual standard error: 16.62 on 9 degrees of freedom
## Multiple R-squared:  0.2579, Adjusted R-squared:  0.01048
## F-statistic: 1.042 on 3 and 9 DF,  p-value: 0.4198
```

```
lm_fit_geno_snp_s <- lm(Observation ~ 0 + `SNP S`, data = tbl_all_data_ascii)
summary(lm_fit_geno_snp_s)
```

```
##
## Call:
## lm(formula = Observation ~ 0 + `SNP S`, data = tbl_all_data_ascii)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2367 -1.2575 -0.8383  0.8238  4.9233
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## `SNP S`S1S1  24.4750     1.7851  13.711 2.46e-07 ***
## `SNP S`S1S2   0.6967     1.0306   0.676    0.516
## `SNP S`S2S2 -22.8700     1.2623 -18.118 2.17e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.525 on 9 degrees of freedom
## Multiple R-squared:  0.9829, Adjusted R-squared:  0.9772
## F-statistic: 172.2 on 3 and 9 DF,  p-value: 2.887e-08
```

From the resulting model fit, it becomes clear, that SNP R has a bad fit whereas SNP S fits the data much better.