57th Annual EAAP Meeting in Antalya, Turkey

Session 3: **Statistical analysis of genetic data – G3.11**

Generalized linear models vs. linear models for the detection of QTL effects on within litter variance

Dörte Wittenburg [1], Volker Guiard, Norbert Reinsch

Forschungsinstitut für die Biologie landwirtschaftlicher Nutztiere FBN
Wilhelm-Stahl-Allee 2, 18196 Dummerstorf, Germany

1 Introduction

Quantitative trait loci (QTL) may not only affect the mean of a trait but also its variability. A special aspect is the variability between multiple measurements of genotyped animals, for example the within litter variance of piglets birth weights. The authors Damgaard et al. (2003) state the view, the within litter variance influences the sow productivity. Evaluating the available data of pigs they assign the within litter standard deviation as maternal trait and detected a maternal additive genetic variance for the within litter standard deviation. A heritability estimate for this trait was 8%, which was significantly larger than zero.
Our present study exemplary adapts the maternal trait by the non-normally distributed sample variance of birth weights within litter and benefits from the adequate approximation of its distribution. To detect QTL effects in the daughter-design a generalized linear model with an identical link function was applied. Adapted test statistics were constructed to evaluate the test problem in terms of statistical power and desired error probability under the null hypothesis H_0: *No QTL with effect on the within litter variance is segregating.* Furthermore we take a look on the estimates of the QTL effect and the QTL position. To compare the advantages of this method with more common tools of statistics, a weighted regression approach was developed, taking a transformed sample variance as observation.

2 Theory

It is assumed, that a population of pigs has two alleles at the QTL denoted by Q and q. Further we look on a fixed number N of sires in our studies, which are drawn by chance of the population. Every sire is mated with n unrelated dams of the population. We pick out one daughter per mating and consider the offspring's birth weight as a multiple measurement. The sample variance of weights at birth within one litter is taken as observation for every daughter. Those daughters, who inherit the QTL allele Q from the presumed heterozygote sire, feature uniformity of birth weights. Daughters inheriting q show an increased variability of birth weights of their offspring. Thus, from the breeder's perspective, the positive effect of the QTL, that means the lower within litter variance, is inherited with the QTL allele Q.

[1] E-mail to wittenburg@fbn-dummerstorf.de

August 16, 2006

The sires may have the marker genotype of kind $m_{l,1}, m_{l,2}$, where $l = 0, 1, \ldots, L$ denotes the marker position on the chromosome. The sire's two marker alleles are denoted by $m_{l,1}$ on the paternal allele and $m_{l,2}$ on the maternal allele, respectively, for every marker position. It is not possible to determine which sire is heterozygote or homozygote at the QTL a priori. After the sires are genotyped, we suppose that all daughters are fully informative, therefore we only need to consider paternal alleles of daughters. The recombination rates are calculated by Haldane's mapping function. We consider intervals flanked by markers of type $m_{l,r}m_{l+1,s}$ with $r, s \in \{1, 2\}$ specifying the marker alleles. The transmission probability for inheriting the QTL allele Q of the heterozygote sire with genotype Qq is determined for every desired position $d = 0, 1, \ldots, D$ (usually in steps of 1 cM) on the chromosome.

2.1 Model and distribution of the data

The phenotypic value $Y_{ijk,d}$ of the piglets within one litter are described by the following model, consisting of independent components with sire $i = 1, \ldots, N$, daughter $j = 1, \ldots, n$ and piglet $k = 1, \ldots, n_{ij}$. Let n_{ij} denote the litter size and capital letters are used for random variables.

$$Y_{ijk,d} = \mu_{ij} + A_{ijk} + G_{ijk} + \mathbf{1}_{\{Q\},ij,d}E_{ijk} + c_* \left(1 - \mathbf{1}_{\{Q\},ij,d}\right) E_{ijk} \tag{1}$$

Whereby μ_{ij} denotes the constant mean value within litter, A_{ijk} declares the normally distributed mendelian sample $\sim N\left(0, \frac{1}{2}\sigma^2_{polygene}\right)$, G_{ijk} is the non-normally distributed additive QTL effect depending on the piglet's genotype and E_{ijk} means the normally distributed random deviation $\sim N\left(0, \sigma^2_e\right)$. The indicator function $\mathbf{1}_{\{Q\},ij,d}$ takes the value 1, if the daughter ij inherits the QTL allele Q at the unknown QTL position d from the sire i. In case of inheriting q the random deviation E_{ijk} of the model (1) is altered by a factor $c_* \in (0, \infty)$.
After transforming the random variables $Y_{ijk,d}$ into variables $X_{ijk,d}$, which only consist of the non-constant terms of the phenotypic value within one litter, the expected value and within litter variance are conditional on $A_\delta = \left\{\mathbf{1}_{\{Q\},ij,d} = \delta\right\}$, $\delta \in \{0, 1\}$

$$\begin{aligned}
\mathbb{E}\left(X_{ijk,d}|A_\delta\right) &= 0 \\
\mathbb{V}\left(X_{ijk,d}|A_1\right) &= \frac{1}{2}\sigma^2_{polygene} + \sigma^2_{QTL} + \sigma^2_e =: \tau^2 + \sigma^2_{QTL} && (2) \\
\mathbb{V}\left(X_{ijk,d}|A_0\right) &= \frac{1}{2}\sigma^2_{polygene} + \sigma^2_{QTL} + c^2_*\sigma^2_e =: \tau^2_* + \sigma^2_{QTL} && (3)
\end{aligned}$$

$\tau^2 = \frac{1}{2}\sigma^2_{polygene} + \sigma^2_e$ summarizes the variance of the normally distributed effects of the phenotypic value under the condition of inheriting the QTL allele Q and $\tau^2_* = \frac{1}{2}\sigma^2_{polygene} + (c_*\sigma_e)^2$ includes the altered residual variance component.

The parameter c^2 is now defined to specify the ratio of the within litter variance when the daughter ij inherits the QTL allele q to the within litter variance in case of inheriting the paternal allele Q from the presumed heterozygote sire

$$c^2 = \frac{\mathbb{E}\left(S^2_{ij,d}|\mathbf{1}_{\{Q\},ij,d} = 0\right)}{\mathbb{E}\left(S^2_{ij,d}|\mathbf{1}_{\{Q\},ij,d} = 1\right)} = \frac{\mathbb{V}(X_{ijk,d}|\mathbf{1}_{\{Q\},ij,d} = 0)}{\mathbb{V}(X_{ijk,d}|\mathbf{1}_{\{Q\},ij,d} = 1)} = \frac{\tau^2_* + \sigma^2_{QTL}}{\tau^2 + \sigma^2_{QTL}} \tag{4}$$

Therefore it appears a test problem, whether the QTL effect on the within litter variance actually

exists, that means the increased sample variance $S^2_{ij,d}$ is conditioned by the inherited paternal QTL allele, or the ratio c^2 is equal to 1.

Because the investigated sample variance consists of non-normally distributed traits, the conditional distribution of $S^2_{ij,d}$ is not χ^2. For example, the distribution of $S^2_{ij,d}$ conditional on $\left\{\mathbf{1}_{\{Q\},ij,d} = 1\right\}$ coincides approximately with a gamma distribution $\Gamma_{\mu,\nu_{ij}}$ with expected value μ and variance $\frac{\mu^2}{\nu_{ij}}$, where

$$\mu = \tau^2 + \sigma^2_{QTL} \quad \text{and} \quad \nu_{ij} = \frac{n_{ij}-1}{2} \frac{\left(\tau^2 + \sigma^2_{QTL}\right)^2}{(\tau^2 + 2\sigma^2_{QTL})\tau^2 + \sigma^4_{QTL}\left(\frac{n_{ij}}{4} + \frac{1}{4} + \frac{1}{n_{ij}}\right)} \tag{5}$$

2.2 Generalized linear models

Let $T_{ij,d}$ denote the random variable, which is realized by the according transmission probability depending on the observed flanking marker alleles per daughter ij at the investigated QTL position $d \in \{0, 1, \ldots, D\}$ and use $t_{ij,d}$ to describe the realized transmission probability per individual ij. In order to fit the sample variance in an adapted manner, a multiplicative model will be considered. Let $\beta_d = (u_{1,d}, \ldots, u_{N,d}, b_{1,d}, \ldots, b_{N,d})^T$ denote the parameter vector consisting of the mean value $u_{i,d}$ per sire and the parameter $b_{i,d}$ describing the relation between the observed traits s^2_{ij} per daughter and the inherited paternal QTL allele expressed by the individual transmission probabilities at the investigated position d. The sample variance is now described by the following model

$$S^2_{ij,d} = \{u_{i,d} + b_{i,d}T_{ij,d}\} \cdot \varepsilon_{ij,d} \tag{6}$$

Where $\varepsilon_{ij,d}$ are independently gamma distributed random variables with expected value 1. The conditional expected value of $S^2_{ij,d}$ given the observed marker alleles is $\mu_{ij,d} = u_{i,d} + b_{i,d}t_{ij,d}$. The expectation is of linear form already, thus the identical link function is used to receive the linear predictor $\eta_{ij,d} = \mu_{ij,d}$.

The application of the generalized linear model theory (McCullagh & Nelder, 1989; Fahrmeir & Kaufmann, 1985) provides several test statistics to check the local null hypothesis $H_{0,d}$: *There exists no QTL on the investigated position d affecting the within litter variance*, which is equivalent to

$$H_{0,d}: \qquad b_{1,d} = \ldots = b_{N,d} = 0 \quad \text{or} \quad \mu_d = \mu^0 \tag{7}$$

To test $H_{0,d}$ the following exemplary test statistic is constructed by utilization of the deviance D

$$\widehat{L}_d = \frac{1}{\widehat{\phi}_{Nn,d}} \left[D(s^2, \widehat{\mu}^0_{Nn}) - D(s^2, \widehat{\mu}_{Nn,d})\right] \tag{8}$$

which coincides with the likelihood-ratio test but includes a consistently estimated dispersion parameter $\widehat{\phi}_{Nn,d}$. To check the global null hypothesis H_0 : *There exists no QTL on the chromosome with effect on the within litter variance* or equivalent $H_0 : b_{1,d} = \ldots = b_{N,d} = 0$ for $d = 0, 1 \ldots, D$

the test statistic $\widehat{L}$ is proposed

$$\widehat{L} = \max_{d \in \{0,1,\ldots,D\}} \widehat{L}_d \tag{9}$$

The permutation test is suggested to construct an adequate threshold value, which leads to the rejection of the null hypothesis (Churchill & Doerge, 1994). The statistic $\widehat{L}$ has been verified by simulations. If the null hypothesis H_0 is rejected, the QTL is estimated on that position, which provides the highest value of the test statistic $\widehat{L}_d$ over all investigated positions $d \in \{0, 1, \ldots, D\}$.

2.3 Weighted regression

To compare the results from section 2.2 with more common tools of statistics, e.g. Haley & Knott (1992), a weighted regression approach was achieved. Applying the logarithm on the sample variances, the data are approximated by a normal distribution. The linear model for the sample variance on a fixed investigated position $d \in \{0, 1, \ldots, D\}$ is defined by

$$\ln S^2_{ij,d} = u_{i,d} + b_{i,d} T_{ij,d} + \varepsilon_{ij,d} \tag{10}$$

Where $\varepsilon_{ij,d}$ are approximately normally distributed random variables with expected value 0 and $T_{ij,d}$ as described above. The parameter $u_{i,d}$ denotes the mean value per family and $b_{i,d}$ declares the linear connection between the observations $\ln s^2_{ij}$ and the inherited paternal QTL allele at the investigated QTL position d. An adapted design matrix is arranged by use of the individual transmission probabilities and the regarding weights are constituted by $\frac{n_{ij}-1}{2}$. The adequate test statistic F_d is constructed to check the local null hypothesis $H_{0,d}$, e.g. Seber (1977). To test the global null hypothesis H_0 the permutation test is used to determine the threshold value and the global test statistic is

$$F = \max_{d \in \{0,1,\ldots,D\}} F_d \tag{11}$$

3 Simulation studies

When genotyping the individuals of the population we assume markers in intervals of 10 cM on a chromosome of length 100 cM ($D = 99$). So we have 11 markers at our disposal ($L = 10$). In the simulation we placed a single QTL at position 25 cM (between the third and fourth marker). We simulated the transformed birth weights $X_{ijk,d}$ with $N = 4$ sires and $n = 200$ daughters per sire. The litter size is poisson distributed with mean value of 10. The factor c_* varies from 1 to 1.4 by 0.1. The gene frequency is presumed to be $\frac{1}{2}$. Covariances between the maternal effects and the direct effects of the piglet were omitted. The flanking marker alleles were drawn by chance corresponding to the inherited paternal QTL allele and the recombination rates.

The simulation was repeated 100 times for every investigated factor c_*. 10,000 permutations of the first dataset were used to determine the chromosomewise threshold value. This critical value was also applied for the following 99 repetitions.

Table 1 summarizes the results of evaluating the simulated data, which are achieved using the test statistics F and $\widehat{L}$. It is obvious, that with increasing factor c_* the empirical power *power_emp* increases. The empirical power was determined as the ratio of significant repetitions of simulations

	test statistic	simulated factor of random deviation c_*				
		1.0	1.1	1.2	1.3	1.4
power_emp	F	0.04	0.50	0.88	0.90	0.90
	$\widehat{L}$	0.03	0.62	0.88	0.90	0.90
mean_detec	F	0.4400	0.3568	0.2552	0.2674	0.2647
	$\widehat{L}$	0.4619	0.3556	0.2727	0.2656	0.2647
variance_detec	F	0.1074	0.0643	0.0158	0.0230	0.0176
	$\widehat{L}$	0.1088	0.0600	0.0280	0.0207	0.0163

Table 1: Summary of simulations (10% of the repetitions with exclusive homozygote sires)

to the total number of repetitions. The empirical power should not exceed the value of 90%, because 10% of the repeated simulations created exclusively homozygote sires at random (all $N = 4$ sires are homozygote) and therefore a QTL effect on the within litter variance can not be detected. Under the null hypothesis H_0 ($c = 1$), the α-level of 5% holds for the verified test statistics. Furthermore in table 1 *mean_detec* declares the average of detected QTL positions and *variance_detec* is the sample variance of estimated positions. With c_*-values of 1.2 and larger, where the power is already very high, both models perform equally well. But with c_* equal to 1.1 ($c = 1.09$) the generalized linear model clearly outperforms the weighted regression approach and provides an extra gain of 12% empirical power. From this summary it is obvious, that the generalized linear model exemplary with use of the test statistic $\widehat{L}$ is, in terms of empirical QTL detection power, superior to the test statistic F resulting from the linear model.
For more details the reader is referred to Wittenburg et al. (2006).

References

Churchill, G. A. & Doerge, R. W. (1994). Empirical threshold values for quantitative trait mapping. *Genetics* 138 963–971.

Damgaard, L. H., Rydhmer, L., Løvendahl, P. & Grandinson, K. (2003). Genetic parameters for within-litter variation in piglet birth weight and change in within-litter variation during suckling. *Journal of Animal Science* 81 604–610.

Fahrmeir, L. & Kaufmann, H. (1985). Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *The Annals of Statistics* 13 342–368.

Haley, C. S. & Knott, S. A. (1992). A simple regression method for mapping quantitative trait loci in lines crosses using flanking markers. *Heredity* 69 315–324.

McCullagh, P. & Nelder, J. A. (1989). *Generalized Linear Models*. London: Chapman and Hall, 2nd ed.

Seber, G. A. F. (1977). *Linear Regression Analysis*. New York: Wiley.

Wittenburg, D., Guiard, V. & Reinsch, N. (2006). Linear and generalized linear models for the detection of QTL effects on within-subject variability. *submitted* .