# A Double Hierarchical Generalized Linear Model For Teat Number In Pigs

*M. Felleki*[1,2] and H. Chalkias[2]

## Introduction

Teat number is an important trait in pig production since each piglet need access to its own teat, due to a complex nursing behavior between the piglets and the sow. The number of teats can vary, but fourteen is most common in the Swedish Yorkshire breed. The main aim of this paper is to understand causes of variation in number of teats. We do also investigate if a linear mixed model with residuals modeled by a generalized linear mixed model, a case of a so called double hierarchical generalized linear model, gives a good description of a data set on teats numbers.

## Material and methods

**Data.** The data was provided by the breeding company Nordic Genetics, and the file was prepared by Nils Lundeheim, Swedish University of Agricultural Sciences.
Data consisted of 47866 animals born between Jan 2007 until April 2009 together with a pedigree file. The variables were number of teats at three weeks of age, litter identity, year-month of birth, herd, sex, litter size (total number born), and birth parity. Analyses were restricted to herds with at least thousand animals born during the period studied. The litter sizes were grouped together for the sizes seven or less and for the sizes twenty or more. Parities four or higher were also collected in one group. Teat number observations below ten and above nineteen (totally 25 observations) were removed from the original data set because some of these observations were wrong.

**Model.** The number of teats were fitted by the animal model $\mathbf{y}=\boldsymbol{\mu}+\mathbf{Xb}+\mathbf{Wp}+\mathbf{Za}+\mathbf{e}$. The vector **b** contained the effects of five independent variables which were herd, sex, birthday, litter size, and parity. The vector **p** was a permanent environmental effects which corresponded to the litter identities, and the vector **a** corresponded to the animal identities in the pedigree. The incidence matrices **X**, **W**, and **Z** were assumed to be known.
The residuals **e** were assumed to be normal distributed with mean zero and variance vector $\boldsymbol{\varphi}$, where $\log(E(\boldsymbol{\varphi}|\boldsymbol{\gamma},\mathbf{s}))=\mathbf{v}+\mathbf{X}\boldsymbol{\beta}+\mathbf{W}\boldsymbol{\gamma}+\mathbf{Qs}$. The vectors $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ were the fixed effects and permanent environmental effects in the residual variance part of the model. The vector **s** corresponded to the sire identities, and the incidence matrix **Q** was assumed to be known. Thus a sire model was used to model the residual variance.

[1]Dalarna University, SE-781 88 Borlänge, Sweden
[2]Swedish University of Agricultural Sciences, P.O.Box 7070, SE-750 07 Uppsala, Sweden

It was assumed that **p**, **a**, **γ**, and **s** were independent and that **p**~N(0,$\boldsymbol{I}\sigma$p), **a**~N(0,**A**$\sigma$a), **γ**~N(0,$\boldsymbol{I}\sigma$γ), and **s**~N(0,**Ã**$\sigma$s). Here **A** was the relationship and **Ã** was the sire relationship matrices.
The model was similar to the model used in Sorensen & Waagepetersen (2003), and Rönnegård et al (2010).
All individuals in a given litter shared the same ancestors, hence the model of the mean, when corrected for fixed effects, was variation between litters and variation between siblings within litters. Similar the model for the residual variance was variation between sires and variation between litters of sires.
We analyzed a mean animal model with residuals having homoscedastic variance (model 1), and the model described above (model 3). A model without the sire random effects was also fitted (model 2).

**Method.** Data was analyzed using the method from double hierarchical generalized linear models presented by Lee and Nelder (2006). We used the ASReml software with steps as described in Rönnegård et al (2010):

1. Fit the mean linear mixed model **y**~N(**μ**+**Xb**+**Wp**+**Za**, $\boldsymbol{I}\sigma^2/\mathbf{w}$), where $\sigma^2$ is an unknown constant, and the weights **w** are initially a vector of 1s.
2. Calculate $\hat{\mathbf{e}}^2/(1-\mathbf{h})$ and $(1-\mathbf{h})/2$, where **h** is the diagonal of the hat matrix.
3. Fit a Gamma generalized linear mixed model with log link and weights $(1-\mathbf{h})/2$ on $\hat{\mathbf{e}}^2/(1-\mathbf{h})$~**v**+**Xβ**+**Wγ**+**Qs**.
4. Update the weights **w** to be used in the mean model. The weights should be the reciprocals of the predicted values from the residual variance model. This is because each $\hat{e}^2/(1-h)$ follows a Gamma((1-h)/2,(1-h)/($2\sigma^2$/w))-distribution. Hence the predicted value comes out with mean $\sigma^2$/w.

Iterate between 1.-4. until convergence. The constant $\sigma^2$ converges to 1.

**Concerns.** A common problem in variance component estimation is slow or no convergence when variance component estimates are close to zero. Attempts to put more structure on the residual variance model by including the animal effect instead of the sire effect did not succeed to converge, probably because the variance component was small.

**Strengths.** The method converges fast when convergence is possible. If a model does not converge, this is also revealing itself on an early stage. Hence it is not time consuming to fit different models. Moreover it is clear when convergence is actually obtained, because $\sigma^2$ converges to 1.

**Time.** For both of the models 2 and 3 the method converged in less than twenty runs of 1.-4., and in time around ten minutes or less on a 16Gb Linux server containing eight 2.66GHz quad core CPUs. Fitting the residual variance model in step 3. was the most time consuming task.

## Results and discussion

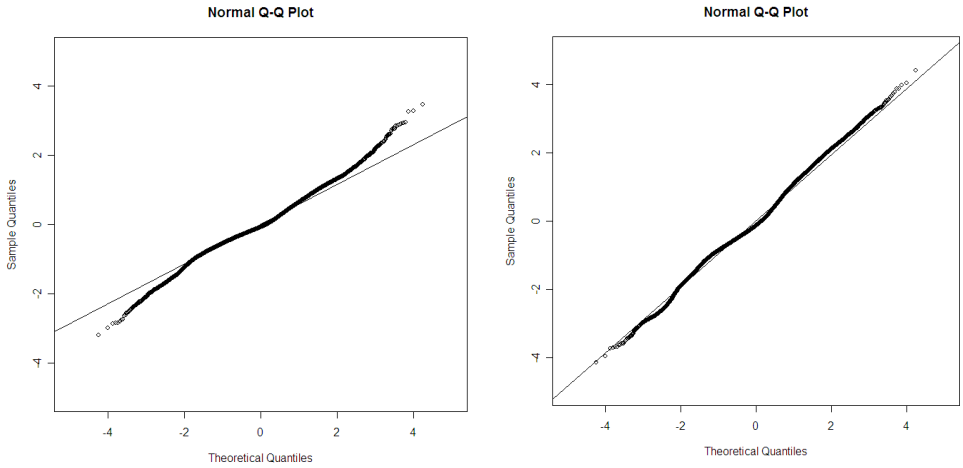
**Variance components.** Estimates of variance components are shown in Table 1. For all of the three mean models the permanent environmental variance components are almost the same, but the animal variance component differs between model 1 and the two other models. This means that a part of the animal effect is explained by heteroscedastic residual variance.

**Table 1: Estimated variance components for the permanent environmental and the animal effects in the mean model and the permanent environmental and the sire effects in the residual variance model.**

| | Mean model | | Residual variance model | |
|---|---|---|---|---|
| Effect | Permanent environmental | Animal | Permanent environmental | Sire |
| 1 | $\sigma p^2=0.03$ | $\sigma a^2=0.35$ | - | - |
| 2 | $\sigma p^2=0.04$ | $\sigma a^2=0.13$ | $\sigma \gamma^2=0.17$ | - |
| 3 | $\sigma p^2=0.04$ | $\sigma a^2=0.14$ | $\sigma \gamma^2=0.14$ | $\sigma s^2=0.03$ |

**Fixed effects.** All estimated fixed effects for the three mean models are found in an interval between -0.2 and 0.5 with only a few outside -0.1 to 0.2. Therefore even though some fixed effects are significant, none beside the constant term adds a big value to the predicted values. The male sex effects (with standard deviations in parenthesis) are 0.049 (±0.008) and 0.045 (±0.007) for the mean models 1 and 3 respectively. For the residual variance models 2 and 3 the estimates are found between 0.6 and 1.3 on a multiplicative scale. Male sex effect for model 3 is 1.087 meaning that the residual variance is 8.7% (7.1%,10.2%) higher for males than for females.

**Model fit evaluation.** QQ-plots for the three models are found in Figure 1. The residuals from model 2 and 3 are heteroscedastic, hence standardization is needed before plotting. To obtain we use the studentized residuals $(\mathbf{w}/\sigma^2)^{½}\hat{\mathbf{e}}/(1-\mathbf{h})^{½}$. For model 1 $\mathbf{w}$ is a vector of ones, and for model 2 and 3 $\sigma^2$ equals one.
It is worth noticing that in some cases, the output from ASReml judge more and more observations as outliers for each run of the algorithm even though the fit is becoming better and better. This is because the residuals in ASReml are not standardized.
The QQ-plots show that model 3 is fitting the data well whereas the residuals from model 1 are not independent identically normal distributed. (Figure 1)




Model 1 Model 3

**Figure 1: QQ-plots of studentized residuals for the models 1 and 3. Model 3 is fitting the data well whereas the residuals from model 1 are not independent identically normal distributed.**

## Conclusion

The number of teats has a hereditary background, whereas the variance of teat number does not seem to be hereditary.

Data is fitted well by the double hierarchical generalized linear model. This model is intuitively easy to understand, fast converging, and it is easy to determine convergence.

## References


Lee, Y., Nelder, J. A. (1996). J R Statist Soc B, 58:619–678.

Lee, Y., Nelder, J. A. (2006). App Stat, 55:139–185.

Lee, Y., Nelder, J. A., Pawitan, Y. (2006). Generalized Linear Models with Random Effects. *Chapman & Hall/CRC*

Rönnegård, L., Felleki, M., Fikse, F., Mulder, H.A., Strandberg, E. (2010) (Submitted)

Sorensen, D., Waagepetersen, R. (2003). *Genet. Res., Camb.*, 82:207-222.