Methods and Statistics of the Social and Behavioral Sciences
Utrecht University, the Netherlands

MSc Thesis *(Pieter Ruben Oosterwijk)*
TITLE: Rasch analysis of chromosomal events as indicators of colorectal tumor severity.
May 2010

Supervisors:
Prof. Dr. *(Paul H. C. Eilers)*
Dr. *(Dave J. Hessen)*

Preferred journal of publication: Statistics in Medicine of Biostatistics
Word count: 4650

Rasch analysis of chromosomal events as indicators of colorectal tumor severity.

Pieter Ruben Oosterwijk

May 19, 2010

Graduate School Utrecht University
Methodology and Statistics in Behavioral and Social Sciences

Authors:
Pieter R. Oosterwijk [1]
Paul H. C. Eilers [2]
Dave. J. Hessen [3]

Abstract

A Single Nucleotide Polymorphism (SNP) set is used to estimate latent tumor severity of colorectal cancer. The Rasch model is applied to a data set of 78 colorectal tumors and 727 cytobands, in order to obtain latent tumor severity estimates and cytoband information estimates. The origin of this project lies in a predominant finding in cancer (including colon cancer) research of the last four decades. Namely, the simultaneous occurrence of cancer with chromosomal abnormalities. The value of tumor severity analysis using Rasch technology lies in the information it can provide to the medical professional with regard to the severity of the tumor using only a biopsy. Hence, this method could be a valuable tool for preoperative staging. The parameter estimates are obtained using maximum likelihood estimation and the appliance of a penalty in the log-likelihood. The penalty technique added to the Rasch model is valuable in its ability to identify model parameter estimates for tumors without chromosomal events on the cytobands. However, too high penalty values lead to non-convergence of model parameter estimates. The interpretation of tumor severity parameter estimates is similar to tumor grades given by medical experts. Besides information on the severity of specific tumors, cytoband estimates give insight to the extent of the relation between chromosomal aberrations on a specific cytoband location and the severity of a colorectal tumor.

Keywords: Single Nucleotide Polymorphism, Colorectal cancer, Rasch Model, Penalized Log-likelyhood, Cytoband

[1] Graduate School Faculty Social Sciences Utrecht, Methods and Statistics
[2] Erasmus Medical Center Rotterdam, Department of Biostatistics
[3] Utrecht University, Faculty Social Sciences, Methods and Statistics

1 Introduction

The distinction between benign tumors (adenomas) and carcinomas with lymph-node metastasis (spreading malignant tumors) is relevant choosing an appropriate treatment for colorectal tumors (Lips et al, 2007)). In order to have adequate preoperative staging, insights into the pathogenesis of colorectal cancer are needed. A predominant finding in the literature of the past decade is the relation between colorectal cancer and chromosomal instability (Goel & Boland, 2010; Grady & Carethers, 2008; Diep, Kleivi, Teixeira & Lindgjrde, 2006). Tumor cells having abberations on hundreds of genes, and structure and copy number of chromosomes (Balmain, Gray & Ponder, 2003; Sheffer, 2009) became of interest for the prediction of the transition from adenoma to carcinoma (Diep et al., 2006; Lips et al., 2007; Lips et al., 2008; Lips et al., 2008b) . This chromosomal or genetic instability is described as aberrant in copy number of genes (both loss and gain) and loss of heterozygosity (LOH) (Goel & Boland, 2010) and is measured using single nucleotide polymorphism (SNP) arrays (Lips et al., 2007) .

In a study of Lips et al. (2007) logistic regression was applied in order to model the progression from adenoma to carcinoma. The use of logistic regression is relatively new (by knowledge of the present author for the first time in Lips et al., (2007) in the modeling of tumor severity. Lips et. al. (2007) recoded normal gene copy number and gain, loss and LOH into a dichotomous "event" "no event" scheme and used aggregated genomic events to the chromosome level to build a quantitative model predicting tumor severity from adenoma to carcinomas with lymph node meta-stasis. This study proposes an extension to logistic regression, still modeling the dichotomous patterns of gene expression. Rasch models have the capacity to create a latent dimension modeling dichotomous scored items (in this case cytobands). This ability makes it a suitable tool for assessing severity of rectal tumors, as the covariation of abberations on the chromosomes is modeled into a latent dimension tumor severity. The Rasch model will give estimates for tumor severity moreover, it is a tool for identifying common implicated regions as biomarkers for the severity of colorectal cancer. Due to lack of knowledge on specific genes and a lack of consensus between studies, a method that can clearly identify cytobands which are key in predicting colorectal cancer can be valuable (Goel & Boland, 2010) .

Before moving on to the application of the Rasch model and its use with respect to SNP arrays, a short explanation of the data seems useful. A SNP refers to the replacement of a base (adenine (A), guanine (G), cytosine(C) and thymine(T)) in a DNA string. This single or point mutation generally concerns the replacement of T by A, G by C or the other way around. The effect of this point mutation is reflected in the copy number of genes and LOH, which henceforth will be referred to as events. Events occur often and are used as predictors of cancer severity (Balmain et al., 2003). As said before, however the significance of the point mutations would be lost if their location on the specific chromosomes could not be identified. The "bar code" used to identify this location is called cytoband. Obtaining results with respect to the copy number of genes LOH or normal copy on different cytobands is usually done when chromatids (the twin pair of a chromosome) are fixed at the centromere during the mitotic metaphase (a phase in cell division) when DNA strands are coiled up. The cytoband is obtained with a staining process with resulting banding patterns. As cytobands are indicators of locations on a chromosome it gives an indication whether the SNP is located at the long end of the chromosome measured from the centromere, called the q arm, or at the short end, called the p arm. As an example the resulting cytoband "2q23.3" can be explained as a location on the second chromosome("2"), measured at the long end from the centromere ("q") and its relative distance from the centromere ("23.3"). This information is important because besides a parameter estimate for each tumor (θ) a parameter estimate (β) will be assigned to each cytoband.

Extending the work of Lips et. al., (2007) this Rasch model is an extension of the logistic regression used to predict tumor severity with events aggregated on the chromosomal level. The parameter estimates θ and β are obtained using multivariate logistic regression creating a single latent predictor tumor severity. A high negative estimate of θ indicates low tumor severity in this population relative to others, high positive θ estimates mean the opposite. The β parameter estimates will give an indication of where on the latent tumor severity scale, that cytoband gives the most information. The reader with experience in modeling SNP arrays in a regression setting is likely to expect problems with over-fit, parameter estimate identification and multi-collinearity as all are present in a regression setting (Eilers, Boer, Ommen & van Houwelingen, 2001). Problems in a regression setting with SNP array data are mostly caused by the independent variables far outnumbering the cases (the regression coefficients of the cytobands far outnumbering the tumors). The problems resulting from this cases to predictors ratio are often solved using penalized regression (Le Cessie & van Houwelingen, 1992; Eilers et al., 2001; Verwij & van Houwelingen, 1994). The Rasch model however, being often criticized for being too restrictive and having under-fit, is not subject to these kinds of problems. Nevertheless, there is a possible other use for the penalty. In this paper the penalty will be used to include tumors without any events in the model, which is useful because θ estimates for tumor severity are assigned to tumors relative to the tumors in the group.

The aim of this article is to obtain the θ and β parameter estimates and infer on their merit in predicting tumor severity and investigate the possible use of the penalty. The validation of this method is done using the information of the study of Lips et. al., (2007). In that study medical experts obtained a biopsy for 78 tumors, and graded the tumors in five severity stages from adenoma to carcinoma with lymph node metastasis. This information will be used to support the Rasch estimation as a tool for tumor grading and preoperative staging. The remainder of this paper will address the Rasch model and its estimation in section 2. Followed by the estimation of the Rasch model and penalty estimation in section 3, followed by the results of the estimation and the discussion in sections 4 and 5 respectively.

2 Data and Rasch model

2.1 Data

The data set used in this paper was previously published by Lips et. al. (2007). It consists of 78 snap frozen colorectal tumors on which four types of gene copy events are measured on 727 cytobands. The four events are three aberration types (gain, loss and LOH) and normal copy number of genes. Illumina SNP arrays were used to identify copy number and LOH of genes using the "beadarraySNP" R-package. For further information the reader is redirected to Lips et. al. (2007). The primary concern of this study is to model deviations from normality (abberations). As a result the dichotomous Rasch model is chosen analyze an event/ no event pattern. Note however that possibilities exist to model nominal data (e.g. normal copy state, gain, loss and LOH) using nominal polytomous Rasch models. For the use of a dichotomous model all the abberations are recoded into the score/event 1 and the normal copy number is recoded into 0. The resulting data set is a matrix which has i = 1,...,n (n = 78) colorectal tumors on its rows and j = 1,...,k (k = 727) cytobands in its columns. Hence every variable X_{ij} has either a 1 or 0 for an event in a tumor on a particular cytoband. From this data set two subsets were analyzed one with all tumors (n = 78) and one with at least one event on a tumor (n = 74). For each tumor a severity grade was available: Tumors with only adenoma tissue (A/A), tumors with adenoma fractions of cases with carcinoma tissue infiltrating at least in the submucosa (A/C), tumor fractions containing a mixture of adenoma and carcinoma tissue (AC/C), tumors with only carcinoma tissue (C/C) and carcinomas with lymph node metastasis (C/C (N+)) [1].

2.2 Rasch model

As said the Rasch model computes a latent trait tumor severity. θ and β are obtained modeling the responses in the variables X_{ij} and the chance on an event π_{ij} displayed in the Rasch model

$$P(X_{ij} = 1|\theta_i, \beta_j) = \pi_{ij} = \frac{e^{\theta_i - \beta_j}}{1 + e^{\theta_i - \beta_j}}. \tag{1}$$

.

Hence, the chance on a chromosomal event π_{ij} is calculated by means of the θ estimate of tumor severity for tumor i and the importance of the information of cytoband j represented by β. The θ parameter estimate is interpreted in this case as the unobserved tumor severity. This trait theoretically ranges from minus infinity to plus infinity. Because an event/abberation is coded as 1 and the presence of abberations is related to the more severe cases of cancer the interpretation of θ is on a scale from least severe to very severe. The β parameter estimate should be interpreted as the value on the latent cancer severity scale where the probability of an event is .5 as can be seen in Equation 1 and will be further addressed later. Hence concerning the β parameter estimate one can say that it gives information for each cytoband were a high value of β should be interpreted as the cytoband giving its optimum of

[1] Like the data the tumor grades are obtained from (Lips et al, 2007)

information on the higher end of the tumor severity scale. For each variable X_{ij} in this matrix it can be shown that the chance on an abberation/event is denoted by π_{ij}, which is displayed in Equation 1 and its corresponding chance on no event(Fischer & Molenaar, 1995), which is displayed as

$$1 - \pi_{ij} = \frac{1}{1 + e^{\theta_i - \beta_j}}. \tag{2}$$

In Equation 1 one can easily see the interpretation of β giving a .5 probability on an event when it is equal to θ, and thus giving its optimum of information.

Under the Rasch model local independence is assumed. Local independence means

$$P(X_{i1} = x(i1), X_{ik} = x_{ik}|\theta_i, \beta_1, ...\beta_k) = \prod_{j=1}^{k} P(X_{ij} = x_{ij}|\theta_i, \beta_j) = \prod_{j=1}^{k} \pi_{ij}^{x_{ij}} (1 - \pi_{ij})^{1-x-ij}. \tag{3}$$

Assuming independence of observations, it follows that the likelihood function is

$$L(\theta_1, ..., \theta_n, \beta_1, ..., \beta_k) = \prod_{i=1}^{n} \prod_{j=1}^{k} \pi_{ij}^{x_{ij}} (1 - \pi_{ij})^{1-x_{ij}}. \tag{4}$$

Substitution from (1) and (2) into (4) and taking the natural logarithm of both sides gives

$$\ell = \sum_{i=1}^{n} \theta_i t_i - \sum_{j=1}^{k} \beta_j s_j - \sum_{i=1}^{n} \sum_{j=1}^{k} ln(1 + e^{\theta_i - \beta_j}). \tag{5}$$

Where t_i is the sum of events over row i and s_j is the sum of events over column j. The estimates of θ and β can be obtained by miximizing the Joint Maximum Likelihood (JML) estimates of Equation 5.

3 Computational statistics

3.1 Newton Raphson estimates

See Appendix II for an R function, executing a Rasch analysis as well as a penalized Rasch analysis. To obtain the parameter estimates β and θ the Newton Raphson algorithm is used. Newton Raphson algorithms are calculated using derivations of the log-likelihood, in this case the JML. The first derivative of the log-likelihood (5) with respect to θ,

$$\frac{\delta \ell}{\delta \theta} = t_i - \sum_{j=1}^{k} \frac{e^{\theta_i - \beta_j}}{1 + e^{\theta_i - \beta_j}}, \tag{6}$$

and the second derivative with respect to θ,

$$\frac{\delta^2 \ell}{\delta \theta^2} = - \sum_{j=1}^{k} \frac{e^{\theta_i - \beta_j}}{(1 + e^{\theta_i - \beta_j})^2}, \tag{7}$$

are implemented in the equation for the Newton Raphson algorithm. This gives the Newton Raphson algorithm for θ, with iterations u

$$\theta_{i(u+1)} = \theta_{i(u)} - \frac{\delta \ell}{\delta \theta_{i(u)}} / \frac{\delta^2 \ell}{\delta \theta_{i(u)}^2}. \tag{8}$$

Newton Raphson iterations for β are obtained by implementing the first derivative of the JML with respect to β,

$$\frac{\delta \ell}{\delta \beta} = -s_j + \sum_{i=1}^{n} \frac{e^{\theta_i - \beta_j}}{1 + e^{\theta_i - \beta_j}}, \tag{9}$$

and the second derivetive of the JML with respect to β,

$$\frac{\delta^2\ell}{\delta\beta^2} = \sum_{i=1}^{n} \frac{-e^{\theta_i-\beta_j}}{(1+e^{\theta_i-\beta_j})^2}, \tag{10}$$

in the equation of the Newton Raphson algorithm. The Newton Raphson algorithm for β with iterations u is then be written as

$$\beta_{j_{u+1}} = \beta_{j(u)} - \frac{\delta\ell}{\delta\beta_{j(u)}} / \frac{\delta^2\ell}{\delta\beta^2_{j(u)}}. \tag{11}$$

In order to obtain JML estimates of β and θ starting values are chosen for θ $(ln(t_i)/(k-t_i))$ and β $(ln(n-s_j)/(s_j))$. The Newton Raphson algorithm for $\theta(8)$ is applied until convergence is reached (Baker & Kim, 2004). A similar process is executed for β (see Equation 11). The parameter estimates β and θ are estimated with updated values for β and θ in each iteration, as the derivatives used in Equation 11 and Equation 8 depend on each other in JML. This is a direct consequence of the different elements of the Newton Raphson algorithm: Equation (6, 7, 9 and 10). A drawback of this basic Rasch model is that it can not model perfect response patterns as they give infinite parameter estimates (Fischer & Molenaar, 1995). Therefore tumors without events or with all events on each tumor, as well as cytobands with all or none events have to be removed for the standard form of Rasch modeling. In this data set, four tumors had no event at any of the cytoband and where therefore removed. In order to obtain converged estimates for β and θ one parameter estimate has to be has to fixed. For this purpose the values of θ are centered around their mean. The consecutive steps for the Newton Raphson algorithm in this case are:

Step 1 Choose Starting values, $\beta_{j,\ldots k,u=0} = ln(n-s_j)/(s_j))$ and $\theta_{i,\ldots n,u=0} = ln(t_i)/(k-t_i)$.

Step 2 Execute one Newton Raphson iteration for θ (see Equation 8).

Step 3 Center θ around its mean.

Step 4 Execute one Newton Raphson iteration for β (see Equation 11).

Step 5 Check for convergence of parameter estimates θ and β

Step 6 Repeat step 2 through 5 until convergeance is reached.

3.2 Penalty estimation

As stated earlier penalized estimation is not necessary in order to obtain estimates of θ and β parameters. However, some aspects of the penalty can be useful. The penalty is implemented in the log-likelihood function before derivation as can be seen below for θ

$$\ell_p = \ell - \lambda \sum_{i=1}^{n} \frac{\theta_i^2}{2}. \tag{12}$$

The derivations of the penalized log-likelihood with respect to θ for the Newton Raphson iterations are easy. Like the penalty for θ a penalty for β can be implemented for β in the stated estimation methods given in the paragraph on Newton Raphson estimates. The function of a penalty constant λ is to discourage high values of the parameter estimate it is penalizing, as with the increase of the parameter estimate directly leading to a decrease if likelihood function as can be seen in 12. This creates an interesting possibility, namely the inclusion of tumors and cytobands with perfect scores.

In the Rasch model this would lead to unidentified parameter estimates because they tend to minus or plus infinity. As stated before the inclusion of tumors with perfect scores can be very important as the benign tumors are expected to have no events and the tumor grade θ is relative too other tumors. The use of this penalty can ensure the possibility to create a norm group for comparison. Different values for penalties with regard to the estimate of θ and β will be implemented and assessed which penalty leads to the lowest value of the AIC (Akaike, 1974).

4 Results

4.1 Rasch Model

The Rasch model is fully identified and gives both β and θ parameter estimates. As can be seen in Figure (1) the tumor grades given by the medical experts correspond with the estimates of θ. One can see a gradual increase in θ as the tumor increases from a benign tumor (adenoma) to a severe carcinoma with lymph-node meta-stasis (”CC (N+)”)

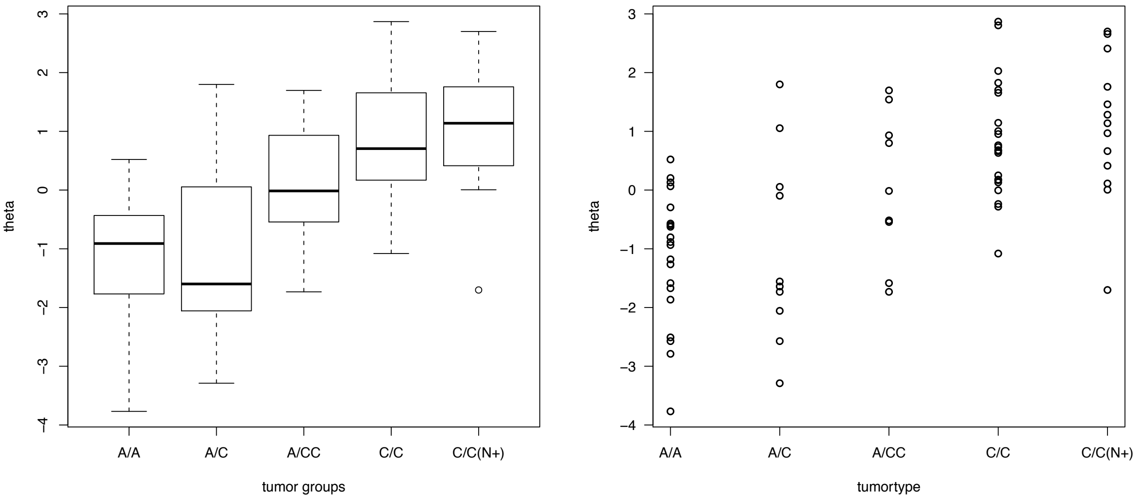

Figure 1: Latent trait θ in relation with observed tumor severity.

Furthermore the test information curve (Baker & Kim, 2004), given in Figure 2, gives a good indication of where on the latent trait scale this Model gives its optimum of information. For the interpretation of this curve and the values of θ one should know that the tumors graded in five groups from adenoma to carcinoma with lymph-node metastasis had a n of respectively 20, 10, 9, 22 and 13. As can be seen in Figure 1 the estimate of θ increased with the increase of the tumor grades given by the medical experts. For a complete overview of the tumors and their severity estimates see Table 1 in Appendix I. Furthermore one can see that if the information of Figure 2 is combined with the count in numbers in the different tumor groups the test information curve is at its highest where the tumor type count is.

With regard to the cytobands it can be seen that many of the adjacent cytoband parameter estimates β have the same value, as expected in research findings (Eilers et al., 2001). In the setting of the Rasch model this means that these items give the same information about the tumor severity. Multi-collinearity is in this case not a problem as the probability of events are functions of the parameter estimates, and not like logistic regression where the estimates predict the tumor severity. The β parameter estimates are summarized in the Appendix I table 2, the top five values for each chromosome and arm are listed when possible. Extreme values are listed and equal values are left out. For an indication of the different β estimates see Figure (2). In this Figure one can see the cytobands implicated the most with colorectal cancer (the higher estimates of θ). Figure 2 shows some slight negative β estimates. However, these are not to be interpreted as having a negative relationship with tumor severity, as one might think baring a regression analysis in mind. These values simply indicate that the optimum of information is somewhat below the mean of θ, since the β parameter estimate should be interpreted as the location where the cytoband gives its optimum of information on the latent tumor severity scale θ.

4.2 Penalized Rasch model

The penalty was successful in estimating tumor parameters without any chromosomal events on the cytobands (for the cases see, Appendix I, table 1 cases 1,2,3 and 24). As the regular Rasch model

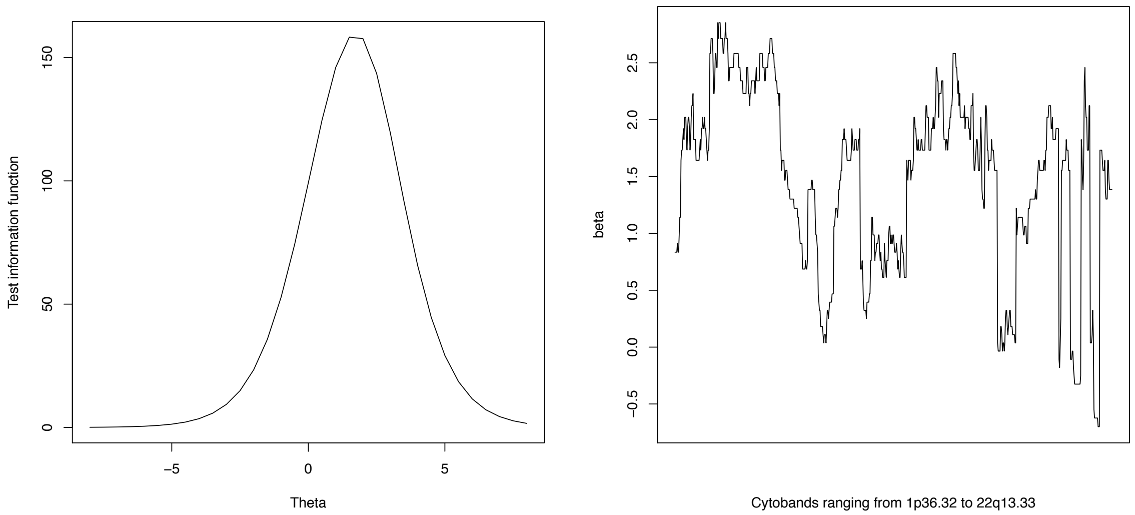

Figure 2: Display of test information, test information curve and β parameter estimates respectively.

cannot identify parameter estimates for these cases sometimes it is possible to add one event so that the parameter estimates are obtained, although it is not recommended (Fischer & Molenaar, 1995) . When comparing the estimate of cases without events for the model with just one event added the difference is apparent. The Rasch model without penalty does not identify the model parameter estimates, the Rasch model with a constant added to each tumor identified the model parameter estimates. However, when the constants (1 event) for the cases which originally had no events are added, extreme estimates are obtained (-38 where the normal scale would range from -4 to 3). The model with a small penalty ($\lambda = .01$ see appendix 1, table 1) obtained measurements the most like the normal regression method. The last model with the largest penalty had parameter estimates most marginalized by the penalty.

Now that the penalty works attempts where made to find the optimal constant to be implemented for the penalty. The AIC for both models was calculated and plotted against penalty values for the θ and β parameter estimates. However the AIC (as can be seen in 3) had no minimum, at the largest possible value of $\lambda = 25$ larger values resulted in convergence problems for the model parameter estimates. Therefore, the AIC was not a useful tool for identifying the appropriate penalty.

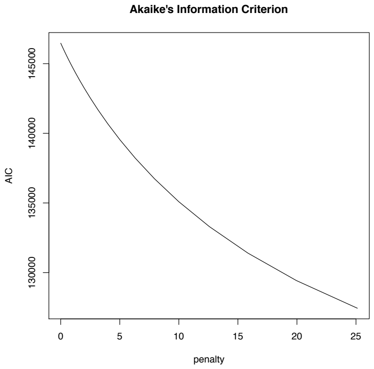

Figure 3: AIC for models penalizing for model parameter estimates

5 Discussion

The dichotomous Rasch model used in this paper produced parameter estimates of β and θ of tumors and cytobands with events > 1. The θ parameter estimates follow same ordering of the observed tumor severity scores (A/A to C/C (N+)). Therefore the model can further be investigated to be eventually used as tumor severity rating tool, which only uses a tumor. Further study is needed with different samples in order to validate the model. Furthermore in order to distinguish adenomas from carcinomas for populations of interest careful construction of norm groups is necessary as this paper only concerns the statistical implementation of the Rasch model on this type of data. Interesting model extensions are also possible. The data used is a display of the presence of a chromosomal abberation, and therefore a dichotomous Rasch model is used. It might be possible that a polytomous Rasch model is fitted on a data set which displays loss, gain or LOH on the cytobands. This was however not the present concern. A final model extension could also be another Item Response model, such as the 2PL model of (Birnbaum, 1968) besides giving an estimate of tumor severity for the tumor and maximum information estimate for the cytoband the IRT model can also present the researcher with a parameter estimate that displays information for each cytoband on the extent it is measuring the latent trait (colorectal tumor severity).

The penalized estimation, which was executed to estimate θ parameter estimates for tumors without events, was successful for that purpose. The resulting tumor severity estimates where in concurrence with theory as the tumors without chromosomal events had the lowest tumor severity estimates. (Appendix 1, Table 1). However the AIC of models with different penalty's did not result in a penalty value with an identified clear minimum. Instead, the penalty decreased with an increase of the penalty value until the size of the penalty caused the parameter estimates to be unidentified. What causes this problem will be further investigated and attempts will be made to find a suitable manner to optimize the model penalty. For the pragmatic reader it might be worthwhile to know that in order to obtain the estimates of θ a very small penalty with minimal influence on β or θ values is enough to identify all tumors severity estimates and produces θ estimates almost equal to normal Rasch analysis.

Further research into penalty estimation of Rasch models could also yield research into the possibility to obtain more exact estimates for a certain cutoff point. As penalty estimation makes model parameter estimates smaller (as it penalizes harder as the estimates are larger) the let penalized β parameter estimates tend to zero. Because θ has to be constrained (usually it would be a mean of zero) one could arbitrarily decide on the constraint. For example if one would have a norm group where 80% of the cases are adenomas and 20% where carcinomas the model constraint for θ could be: $\theta - .8\theta$. When the β parameter estimates tent to zero it might be possible that it optimizes the information around the chosen constraint of θ. For this to work it is necessary to find a way to optimize θ.

References

Akaike, H. (1974). A new look at the statistical model identification. system identification and time-series analysis. *IEEE Trans. Automatic Control*, *19*, 716 - 723.

Baker, F. B., & Kim, S. (2004). *Item response theory: Parameter estimation techniques.* Taylor & Francis Group.

Balmain, A., Gray, J., & Ponder, B. (2003). The genetics and genomics of cancer. *Nature Genetics*, *33*, 238-244.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. in f. m. lord & m. r. novick (eds.). *Statistical theories of mental test scores*, 379-479.

Diep, C., Kleivi, F., K.and Ribeiro, Teixeira, M., Lindgjarde, O., & Lothe, R. (2006). The order of genetic events associated with colorectal cancer progression inferred from meta-analysis of copy number changes. *Genes Chromosomes and Cancer*, *45*, 31-41.

Eilers, P. H. C., Boer, J. M., Ommen, G. J. van, & Houwelingen, J. C. van. (2001). Classification of microarray data with penalized logistic regression. In M. L. Bittner, Y. Chen, A. N. Dorsel, & E. R. Dougherty (Eds.), (Vol. 4266, p. 187-198). SPIE.

Fischer, G. H., & Molenaar, I. W. (Eds.). (1995). *Rasch models: Foundations, recent developments, and applications.* Springer-Verlag.

Goel, A., & Boland, C. R. (2010). Recent insihgts into the pathogenesis of colorectal cancer. *Corrent Opinion in Gastroenterology*, *26*, 47-52.

Grady, W., & Carethers, J. (2008). Genomic and epigenetic instability in colorectal cancer pathogenesis. *Gastroenterology*, *135*, 1079-1099.

Le Cessie, S., & Van Houwelingen, J. (1992). Ridge estimators in logistic regression. *Applied Statistics*, *41*, 191201.

Lips, E., Eijk, R. van, Graaf, E. de, Oosting, J., Miranda, N. de, Karsten, T., et al. (2008). Integrating chromosomal aberrations and gene expression profiles to dissect rectal tumorigenesis. *BMC Cancer*, *8*.

Lips, E., Van Eijk, R., De Graaf, E., Doornebosch, P., De Miranda, J., N.F.C.C.and Oosting, Karsten, T., et al. (2008). Progression and tumor heterogeneity analysis in early rectal cancer. *Clinical Cancer Research*, *14*, 772-781.

Lips, E. H., Graaf, E. J. de, Tollenaar, R. A. E. M., Eijk, R. van, Oosting, J., Szuhai, K., et al. (2007). Single nucleotide polymorphism array analysis of chromosomal instability patterns discriminates rectal adenomas from carcinomas. *Journal of Pathology*, *212*, 269-277.

Sheffer, M., Bacolod, M., Zuk, O., Giardina, H., S.F.AND Pincas, Barany, F., Paty, P., et al. (2009). Association of survival and disease progression with chromosomal instability: A genomic exploration of colorectal cancer. *Proceedings of the National Academy of Sciences of the United States of America*, *106*, 7131-7136.

Verweij, P., & Van Houwelingen, J. (1994). Penalized likelihood in cox regression. *Statistics in Medicine*, *13*, 24272436.

Appendix I

Table 1: Tumor parameter θ for different tumor (λ) and cytoband penalty's (κ)

Tumor number	Rasch model			Penalized Rasch	
	Tumor Grade	No penalty	event +1	$\lambda = 0$, $\kappa = 15$	$\lambda = .1$ $\kappa = 0$
1	AA	*	-38.021	-38.021	-5.656
2	AA	*	-38.021	-38.021	-5.656
3	AA	*	-38.021	-38.021	-5.656
4	AA	-3.768	-1.780	-1.780	-3.396
5	AA	-2.789	-0.798	-0.798	-2.463
6	AA	-2.572	-0.581	-0.581	-2.251
7	AA	-2.508	-0.517	-0.517	-2.188
8	AA	-1.866	0.130	0.130	-1.552
9	AA	-1.671	0.326	0.326	-1.358
10	AA	-1.584	0.414	0.414	-1.271
11	AA	-1.264	0.737	0.737	-0.952
12	AA	-1.180	0.823	0.823	-0.867
13	AA	-0.936	1.071	1.071	-0.623
14	AA	-0.885	1.122	1.122	-0.572
15	AA	-0.804	1.205	1.205	-0.490
16	AA	-0.625	1.387	1.387	-0.311
17	AA	-0.597	1.415	1.415	-0.283
18	AA	-0.570	1.443	1.443	-0.256
19	AA	-0.295	1.725	1.725	0.021
20	AA	0.064	2.093	2.093	0.381
21	AA	0.131	2.162	2.162	0.447
22	AA	0.205	2.238	2.238	0.521
23	AA	0.521	2.566	2.566	0.838
24	AC	*	-38.021	-38.021	-5.656
25	AC	-3.289	-1.300	-1.300	-2.947
26	AC	-2.572	-0.581	-0.581	-2.251
27	AC	-2.057	-0.062	-0.062	-1.742
28	AC	-1.733	0.264	0.264	-1.419
29	AC	-1.641	0.356	0.356	-1.328
30	AC	-1.556	0.442	0.442	-1.244
31	AC	-0.097	1.928	1.928	0.219
32	AC	0.055	2.083	2.083	0.371
33	AC	1.054	3.125	3.125	1.365
34	AC	1.799	3.933	3.933	2.087
35	ACC	-1.733	0.264	0.264	-1.419
36	ACC	-1.584	0.414	0.414	-1.271
37	ACC	-0.543	1.470	1.470	-0.229
38	ACC	-0.517	1.498	1.498	-0.202
39	ACC	-0.015	2.012	2.012	0.301
40	ACC	0.800	2.858	2.858	1.115

Note:* unidentified

Continues on the next page

Table 2: Tumor parameter θ for different tumor (λ) and cytoband penalty's (κ)

Tumor number	Rasch model			Penalized Rasch	
	Tumor Grade	No penalty	event +1	$\lambda = 0$, $\kappa = 15$	$\lambda = 14$ $\kappa = 0$
41	ACC	0.932	2.997	2.997	1.245
42	ACC	1.541	3.649	3.649	1.841
43	ACC	1.697	3.820	3.820	1.990
44	CC	-0.237	1.783	1.783	0.078
45	CC	-0.005	2.022	2.022	0.312
46	CC	0.168	2.200	2.200	0.485
47	CC	0.178	2.210	2.210	0.494
48	CC	0.679	2.731	2.731	0.994
49	CC+	2.409	4.636	4.636	2.651
50	CC	-1.081	0.923	0.923	-0.768
51	CC	-0.283	1.736	1.736	0.032
52	CC	0.131	2.162	2.162	0.447
53	CC	0.250	2.285	2.285	0.567
54	CC	0.633	2.682	2.682	0.948
55	CC	0.664	2.715	2.715	0.979
56	CC	0.733	2.787	2.787	1.048
57	CC	0.763	2.818	2.818	1.078
58	CC	0.954	3.019	3.019	1.267
59	CC	1.004	3.072	3.072	1.316
60	CC	1.145	3.222	3.222	1.455
61	CC	1.656	3.775	3.775	1.951
62	CC	1.704	3.827	3.827	1.997
63	CC	1.827	3.964	3.964	2.113
64	CC	2.028	4.190	4.190	2.302
65	CC	2.805	5.127	5.127	2.996
66	CC	2.868	5.209	5.209	3.049
67	CC+	-1.701	0.296	0.296	-1.388
68	CC+	0.005	2.032	2.032	0.321
69	CC+	0.112	2.142	2.142	0.428
70	CC+	0.414	2.455	2.455	0.731
71	CC+	0.664	2.715	2.715	0.979
72	CC+	0.968	3.035	3.035	1.281
73	CC+	1.138	3.215	3.215	1.448
74	CC+	1.283	3.370	3.370	1.590
75	CC+	1.460	3.561	3.561	1.762
76	CC+	1.758	3.887	3.887	2.049
77	CC+	2.658	4.940	4.940	2.870
78	CC+	2.700	4.993	4.993	2.906

Note:* unidentified

Table 3: Cytoband parameter β of Rasch model without penalty

Cytoband	θ	Cytoband	θ	Cytoband	θ
1p12.	2.23	7p15.2	0.32	14q32.2	1.30
1p13.2	2.12	7p14.1	0.25	14q11.2	1.22
1p32.2	2.02	7q21.11	1.14	14q32.13	1.22
1p33.	1.92	7q21.13	0.99	14q13.2	1.14
1p34.1	1.83	7q22.1	0.76	14q13.1	1.06
1q42.2	1.64	7q22.2	0.83	15q26.2	1.73
1q43.	1.73	7q31.1	0.91	15q21.3	1.64
1q42.12	1.83	8p11.1	1.06	15q21.2	1.55
1q44.	1.92	8p12.	0.99	15q21.1	1.47
2p16.3	2.85	8p23.3	0.91	15q14.	1.38
2p25.1	2.71	8p23.2	0.76	16p13.12	2.12
2p25.3	2.58	8p23.1	0.69	16p13.3	2.02
2p22.1	2.46	8q11.21	0.99	16q12.1	1.92
2p23.1	2.34	8q11.1	0.91	16q23.3	1.92
2q13.	2.85	8q13.2	0.83	16q21.	1.83
2q12.2	2.71	8q21.3	0.76	17p13.2	-0.11
2q11.2	2.58	8q21.2	0.69	17p13.1	-0.18
2q22.1	2.46	9p24.3	1.64	17p12.	0.04
2q21.2	2.34	9p24.2	1.55	17p11.2	0.25
2q37.1	2.23	9p24.1	1.47	17q11.1	1.55
3p25.3	2.46	9q12.	2.02	17q12.	1.64
3p22.2	2.34	9q13.	2.02	17q23.2	1.83
3p26.3	2.23	9q21.11	1.92	17q23.3	1.73
3p24.1	2.12	9q21.13	1.83	17q24.3	1.55
3q23.	2.71	9q21.2	1.73	18p11.32	-0.11
3q11.2	2.58	10p11.21	2.12	18p11.22	-0.04
3q13.12	2.46	10p13.	1.92	18q11.2	-0.18
3q13.31	2.34	10p12.31	1.92	18q12.1	-0.25
3q27.1	2.23	10p15.3	1.73	18q12.2	-0.33
3q27.2	2.23	10q11.1	2.46	18q23.	-0.25
3q27.3	2.23	10q11.22	2.34	19p13.3	1.83
3q29.	2.23	10q21.2	2.23	19p13.2	1.73
3q28.	2.12	10q26.13	2.12	19p13.12	1.47
4p16.3	1.73	10q21.1	2.02	19p13.11	1.38
4p15.33	1.64	11p15.5	2.58	19p12.	1.64
4p16.1	1.55	11p14.3	2.46	19q12.	2.34
4p15.1	1.47	11p14.1	2.34	19q13.11	2.46
4q11.	1.55	11p13.	2.23	19q13.12	2.12
5p14.2	1.47	11p11.2	2.12	19q13.2	2.02
5p15.33	1.38	11q25.	2.23	19q13.42	2.12
5q11.2	1.14	11q13.5	2.12	20p13.	0.04
5q12.1	0.99	11q11.	2.02	20p12.1	0.11
5q12.3	0.91	11q14.1	1.92	20p11.23	0.32
6p11.2	1.92	11q23.2	1.83	20p11.21	0.11
6p12.3	1.83	12q11.	2.02	20q11.21	-0.55
6p21.2	1.55	12p12.3	1.83	20q11.22	-0.62
6p21.32	1.47	12p13.33	1.73	20q13.31	-0.70
6p22.1	1.38	12p11.1	1.73	21q11.2	1.73
6p24.3	1.22	12p13.2	1.64	21q22.11	1.64
6p25.1	1.14	12p13.32	1.55	21q22.12	1.55
6p25.3	1.06	12q14.2	2.12	21q22.3	1.64
6q22.1	1.92	12q15.	2.02	22q11.1	1.38
6q11.1	1.83	12q12.	1.83	22q11.21	1.30
6q13.	1.73	12q21.1	1.73	22q12.1	1.64
6q14.1	1.64	12q21.32	1.64	22q12.3	1.47
7p21.2	0.76	13q12.12	-0.04	22q13.1	1.38
7p22.3	0.69	13q12.11	0.04		
7p21.1	0.61	13q13.3	0.11		
7p11.2	0.47	13q13.1	0.18		
7p15.3	0.40	13q31.1	0.32		

Appendix II

R-code, consisting of a function used to run a Rasch analysis. Kappa is the cytoband penalty, lambda the tumor penalty, X is the data matrix with cytobands in the columns and tumors in the rows.

```
Rasch_pro = function(X, kappa., lambda.){

    t.  = as.matrix(apply(X, 1, sum))
    s.  = as.matrix(apply(X, 2, sum))
    n = nrow(t.)
    k = nrow(s.)

    theta.  = 1 * (log((t.+1)/(k - t.)))                      # setting start values for theta
    theta.  = theta.  - mean(theta.)                          # constraining one parameter
    beta.  = log((n - s.)/(s.))                               # setting start values for beta
    convergeance = 0

# for loop in order to find converged beta and theta parameters
    for (i in 1:40){

# Obtaining theta
    thet.m1 = matrix(theta., nrow = n, ncol = k, byrow = FALSE)
    bet.m2 = matrix(beta., nrow = n, ncol = k, byrow = TRUE)
    exp.m = exp(thet.m1 - bet.m2)

    inp = exp.m/(1 + exp.m)                                   # numerator N.R. algorithm
    num = t.  - apply(inp, 1, sum) - (theta.  * lambda.)      # First derivative for theta
    inp2 = exp.m/((1 + exp.m)2̂)                               # denominator N.R. algorithm
    denom = - apply(inp, 1, sum) - lambda.                    # Second derivative for theta
    theta.  = theta.  - (num / denom)                         # New theta estimate
    theta.  = theta.  - mean(theta.)                          # Fixing parameter for convergeance
    dif1 = thet.m1[,1] - theta.                               # convergeance

# Obtaining beta
    thet.m1 = matrix(theta., nrow = n, ncol = k, byrow = FALSE)
    bet.m2 = matrix(beta., nrow = n, ncol = k, byrow = TRUE)
    exp.m = exp(thet.m1 - bet.m2)

    inp = exp.m/(1 + exp.m)                                   # numerator N.R. algorithm
    num = (- s.  + apply(inp, 2, sum)) - (beta.  * kappa.)    # First derivative for beta
    inp2 = -exp.m /(1 + exp.m)2̂                               # denominator N.R. algorithm
    denom = apply(inp2, 2, sum) - kappa.                      # Second derivative for beta
    beta.  = beta.  - (num / denom)                           # New beta estimate
    dif2 = bet.m2[1,] - beta.                                 # for convergeance estimate

    dif = (sum(abs(dif1)))+ (sum(abs(dif2)))                  # Convergeance
    convergeance = c(convergeance, dif)
    if (dif < 1e-3) break
    }
# AIC creating the AIC
    penalbeta = kappa.* sum((beta.2̂)/2))                      # Penalty for beta
    penaltheta = lambda.* sum((theta.2̂)/2)                    # Penalty for theta
loglik = sum(theta.  * t.)  - sum(beta.  * s.)  - sum(log(1 + exp(exp.m))) - penalbeta - penaltheta
                                                              # Loglikelihood of JML of Rasch model.
    AIC. = 2 * (2*(k-1)) - 2 * loglik

return(list(theta.  = theta., beta.  = beta., convergeance = convergeance, AIC. = AIC.))
}
```