Studying DNA Damage and Repair When Exposed to chemicals

Yunran Chen

Abstract

We study the relation between DNA damage and dose of chemical exposures and how the relation may vary across replications and be impacted by repair time. In particular, we measure DNA damage with Olive tail moment and apply a log linear model including dose, repair time, replication and their interaction terms. We find that variation across replications does not exist. Only when no DNA repair happens, a statistically significantly increasing trend in DNA damage with the dose of chemical exposure. Also, only when no DNA repair happens, first dose level leads to more DNA damage compared to the case with no chemical exposure. We also illustrate why Olive tail moment is the best surrogate measurement for DNA damage and further extend our model to multiple cell lines analysis.

1 Introduction

DNA damage may happen during normal metabolic activities or when cells are exposed to environmental chemicals. A DNA repair process always response to these damages rapidly but repair time varies across different damages. A recent study applied single cell gel electrophoresis (comet assay) to record the amount of DNA damage when exposed to the chemical H_2O_2 at different dose levels within different repair time. By this technique, each cell can be represented by a comet, whose tail's size is positively related to the level of DNA damage. Comet assay for 50 cells are recorded simultaneously under a given experiment condition, where the dose level is chosen from 0,5,20,50,100, and repair time is chosen from 0,60,90. For nearly all experimental conditions, 2 replications are conducted. The resulting dataset contains 1400 records, including dose levels, repair time, replication number (1,2), size of comet area and surrogate measurements for DNA damage, such as proportion of DNA in the tail of the comet(TDNA), tail moment (Ttmom, measure of size of tail), Olive tail moment(Otmom), and tail length(tailL). We also include a DNA damage measurement defined as the ratio of the proportion of DNA in the tail versus that in the head (denoted THDNA). Notice the replication number is nested within experimental conditions.

Our main goal is to explore how different chemical exposures relate to DNA damages, and how the repair time and different replications influence such relation. For a future analysis, we are interested in choosing the single most reliable measure for DNA damage, and extending the analysis to study the DNA damage and repair for multiple cell lines. The main challenge is to measure the DNA damage since multiple surrogate measurements are available. We consider selecting one single measurement for DNA damage instead of incorporating multiple measurements.

2 Materials & Methods

We consider applying a generalized linear model $E(Y|X) = \mu = g^{-1}(X\beta)$ to explore the relation between dose of exposure and DNA damage, where Y represents DNA damage, and predictor X include the potential factors on DNA damage. The form of link function depends on the choice of response variables. Specifically, we choose Olive tail moment (Otmom) as the response variable and apply logarithm transformation due to its long tail behavior. We include dose (Dose), repair time (RTime), replication (rep), and also include comet area (CArea) to control its potential effects. We introduce interaction between dose and repair time to capture the impact of repair time on the relation between dose and DNA damage. Since replication number is nested within experimental conditions, to explore the variation across replications, we need to include interactions between replication and every term related to the experimental conditions. We transform dose,

repair time and replication number into dummy variables and scale comet area ($CArea^*$). We consider a full model as shown in equation (1).

$$\begin{aligned} \log(Otmom_i) = \quad & \beta_0 + \sum_{j=1}^{4} \beta_{1,j} I(Dose_i = j) + \sum_{k=1}^{2} \beta_{2,k} I(RTime_i = k) \\ & + \sum_{j=1}^{4} \sum_{k=1}^{2} \beta_{3,jk} I(Dose_i = j, RTime_i = k) \\ & + \beta_{4,r} I(rep_i = 2) + \sum_{j=1}^{4} \beta_{5,jr} I(Dose_i = j, rep_i = 2) + \sum_{k=1}^{2} \beta_{6,kr} I(RTime_i = k, rep_i = 2) \\ & + \sum_{j=1}^{4} \sum_{k=1}^{2} \beta_{7,jkr} I(Dose_i = j, RTime_i = k, rep_i = 2) \\ & + \beta_8 CArea_i^* + \epsilon_i \qquad (1) \end{aligned}$$

where $\epsilon_i \sim N(0, \sigma^2)(iid)$; i, j, k represent individual cell $i \in \{1, ..., 1400\}$, dose level j and repair time level k respectively. $j = 1, ..., 4$ correspond to dose level $5, 20, 50, 100$ respectively. $k = 1, 2$ correspond to repair time $60, 90$ respectively. [1]

This model can capture the relation between dose of exposure and DNA damage and allow for the variation across replicates as well as different repair time. The confidence intervals of coefficients and ANOVA tests on group of coefficients based on the model can help to quantify impact of dose of exposure and infer how repair time and replications influence such impacts.

Since there is no significant heterogeneity across different replications, our final model is as follows:

$$\begin{aligned} \log(Otmom_i) = \quad & \beta_0 + \sum_{j=1}^{4} \beta_{1,j} I(Dose_i = j) + \sum_{k=1}^{2} \beta_{2,k} I(RTime_i = k) \\ & + \sum_{j=1}^{4} \sum_{k=1}^{2} \beta_{3,jk} I(Dose_i = j, RTime_i = k) + \beta_8 CArea_i^* + \epsilon_i \qquad (2) \end{aligned}$$

3 Results

3.1 Exploratory Data Analysis

Figure 1 shows the correlation among the variables. DNA damage measurements are positively related to dose level, suggesting there may exist an increasing trend in DNA damage with the dose of chemical exposure. The size of comet area ($CArea$) is related to dose, repair time and DNA damage, suggesting including $CArea$ as a predictor to control its effect on DNA damage. Unlike other surrogate measurements for DNA damage, the size of comet tail $tailL$ shows a little correlation to dose and a relative weak correlation to other DNA damage measurements, indicating that $tailL$ may not capture the variation of DNA damage under different dose level (not sensitive to variation of DNA damage) and is not consistent with other DNA damage measurements. Based on the prior knowledge, we know $tailL$ is subject to outliers and is hard to measure. Therefore, we conclude $tailL$ is not a suitable surrogate measurement for DNA damage.

Figure 2 shows distributions of different DNA damage measurements under different dose levels and repair time levels in the first replication. Under all four DNA damage measures, we can observe similar patterns. If no repair happens, the DNA damage increases as dose of chemical exposure increases. However, such increasing trend becomes non-significant if DNA repair happens. Compare across different panels, the relative length of box versus length of line for the first panels is smaller than the others, indicating the distribution of $\log(Otmom)$ is less spread out compare to that of other measurements. If we consider a log linear regression, we may expect choosing $Otmom$ as response variable may result in the best model fitting. Figure 3 shows distribution of DNA damage under the same experimental condition but a different replication. We observe similar patterns as in Figure 2, suggesting there may not exist variation across replications.

3.2 Model Estimation and Inference

For the full model, except for the term related to replication, all the coefficients are significantly non-zero as suggested by t-test. We conducted an ANOVA test with H_0 : all the coefficients related to rep ($\beta's$ includes subscript r) equals to 0 versus H_1 : at least one coefficient does not equals to 0. We accept the null hypothesis since the p-value is 0.1575 (F-stat: 1.3883, degree of freedom: 13). We are 95% confident that

[1] r is a notation indicating the coefficient is related to the second replication under the same experimental condition.

no significant variation exists across replicates. We exclude replication in our model for the further analysis and obtain the estimations of the final model shown in Table 5.

Under the final model, we first conduct an ANOVA test to check whether the interactions terms between dose and repair time are all equal to 0. We obtain a p-value$< 2.2e-16$(F-stat:18.982,degree of freedom:8), suggesting the inference of the potential effect of dose level is influenced by repair time. Figure 4 shows confidence intervals of $\beta_{1,j}+\beta_{3,jk}$, representing the unit increase of $\log(\text{Otmom})$ under dose level j compared to 0 dose of exposure at a given repair time k. In no DNA repair happens, the confidence interval is at a increasing order as dose of exposure increase, suggesting a increasing trend in DNA damage exists as the dose of exposure increases. However, when the repair time is 60 or 90, we cannot observe such an increasing trend. The first lines in each panel represent the unit increase of $\log(\text{Otmom})$ under the first dose level compared to no exposure. Only when no repair happens, the confident interval lies above 0, suggesting the DNA damage increase significantly at the first dose level compared to no exposure only when no repair happens. Specifically, when no repair happens, the size of tail is expected to increase by 17.78% to 81.10% at first dose level compared to no exposure to H_2O_2.

A good measure for DNA damage is expected to be interpretable, stable (not subject to outliers), and can be well modeled given the dataset. From the exploratory analysis and prior information, we first exclude tailL. We compare the rest of surrogate measurements by the performance (measured by R^2,BIC) of log linear model (see Table 6). A greater value of R^2 indicates the variation in the response can be better explained by the predictors. A smaller BIC suggests the model has stronger support from the data. Otmom is best surrogate DNA damage measurement supported by the data under the log linear model assumption[2].

The model shown in equation (2) can be easily extended to model multiple different cell lines by introducing random effects for different cell lines. We can consider a mixed effect model shown in equation (3) with each cell line as a group, each cell as an individual. $\beta_{1,j}+\beta_{3,jk}$ represents the impact of dose level j and repair time k for a typical cell line compared to no exposure and no repair happen. $\beta_{1,j}+\beta_{3,jk}+b_{1,jc}+b_{3,jkc}$ represents the impact for cell line c.

$$\begin{aligned}\log(\text{Otmom}_{ic}) = \quad & \beta_0 + \textstyle\sum_{j=1}^{4} \beta_{1,j} I(\text{Dose}_{ic} = j) + \sum_{k=1}^{2} \beta_{2,k} I(\text{RTime}_{ic} = k) \\ & + \textstyle\sum_{j=1}^{4} \sum_{k=1}^{2} \beta_{3,jk} I(\text{Dose}_{ic} = j, \text{RTime}_{ic} = k) + \beta_8 \text{CArea}_i^* \\ & + b_{0,c} + \textstyle\sum_{j=1}^{4} b_{1,jc} I(\text{Dose}_{ic} = j) + \sum_{k=1}^{2} b_{2,kc} I(\text{RTime}_{ic} = k) \\ & + \textstyle\sum_{j=1}^{4} \sum_{k=1}^{2} b_{3,jkc} I(\text{Dose}_{ic} = j, \text{RTime}_{ic} = k) + \epsilon_{ic} \qquad (3)\end{aligned}$$

where c represent cell line c, $b_{\cdot,c} \sim MVN(0, D)(i.i.d), \epsilon \sim N(0, \sigma^2)(iid)$ and $b_{\cdot,c}$.

4 Discussion

For the DNA damage, we consider selecting one measurement for DNA damage instead of incorporating multiple measurements, since a single measurement is easy to interpret and to model. However, if a sets of surrogate measurements capture small but different parts of information, incorporating multiple measurements may be a better idea. The choice of single measurement is related to whether model assumption for different measurements can best fit the data. However, we may prefer a measurement that can be well explained by a simple model.

A Bayesian model may bring more flexibility to the inferences. For example, we can estimate the posterior probability of increasing trend in DNA damage with dose of H_2O_2 ($Pr(\beta_{1,1} + \beta_{3,1k} < ... < \beta_{1,4} + \beta_{3,4k})$), which is hard to achieve in a frequentist' framework.

In addition, Bayesian models show more flexibility on hierarchical structure modeling and dealing with measurement errors. For our analysis, we assume the dataset is ideal where each cell gets the exact dose indicated by the experimental dose. But in practice, there exists heterogeneity among cells in their level of exposure even at the same experimental dose. If we have prior on the measurement error on the dose level, we can treat the dose level as a continuous variable and introduce a latent variable $\text{Dose}_{true} \sim N(\text{Dose}_{obs}, s^2)$ and consider regression on Dose_{true} instead of Dose_{obs}.

[2]We can construct different models for different DNA damage measurements and conduct a model comparison. For TDNA/100, we also consider a beta regression, but it still has a small pseudo-R squared value.

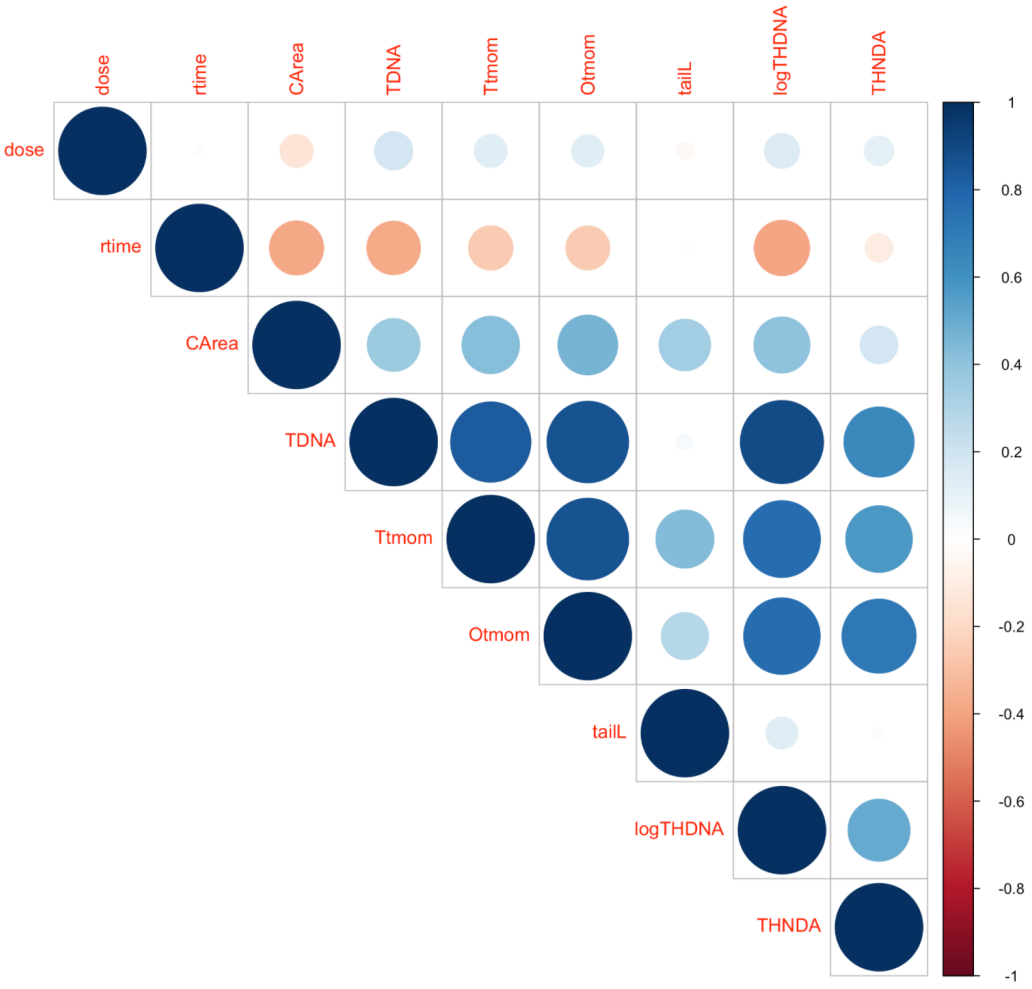

Figure 1: Correlation Between Dose, Repair Time and DNA Damage Measurements.

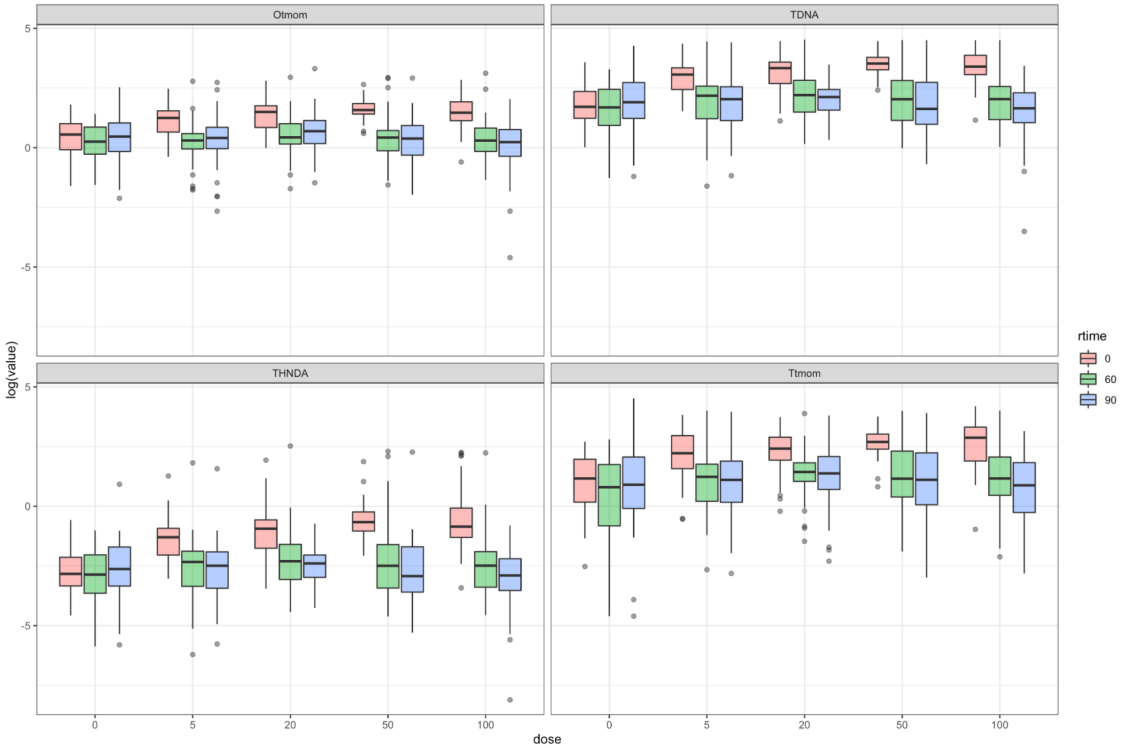

Figure 2: Different DNA damage measurements under different dose levels and repair time levels in the first replication.

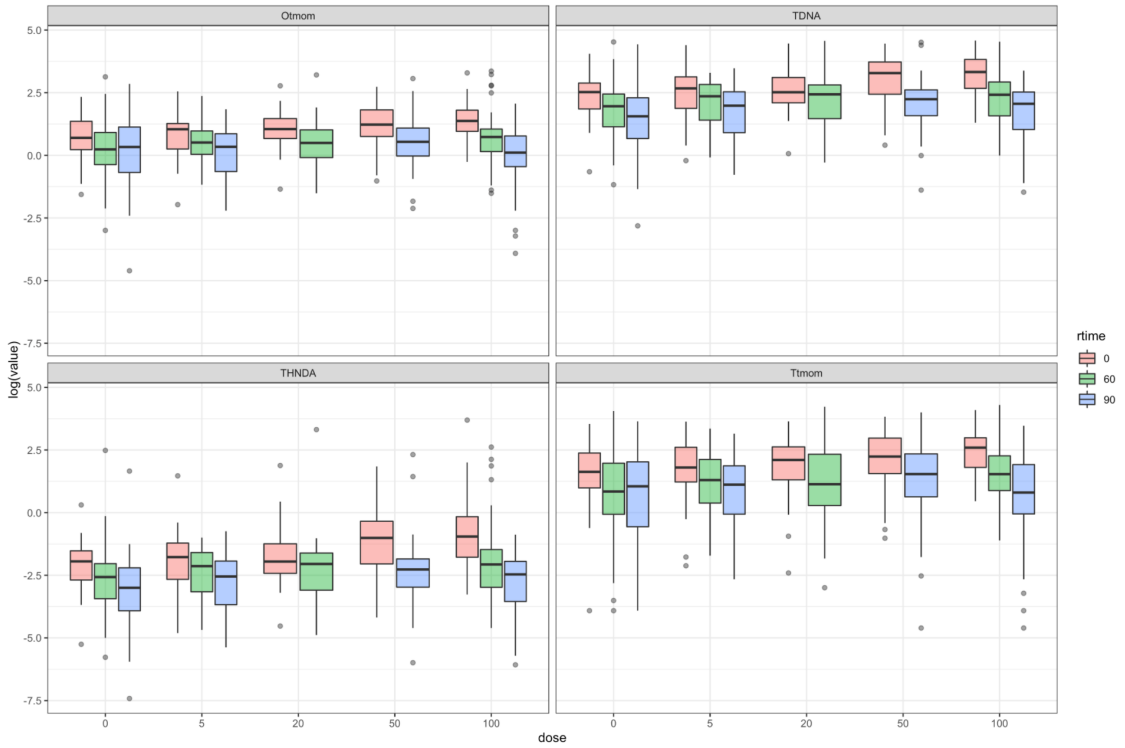

Figure 3: Different DNA damage measurements under different dose levels and repair time levels in the second replication.

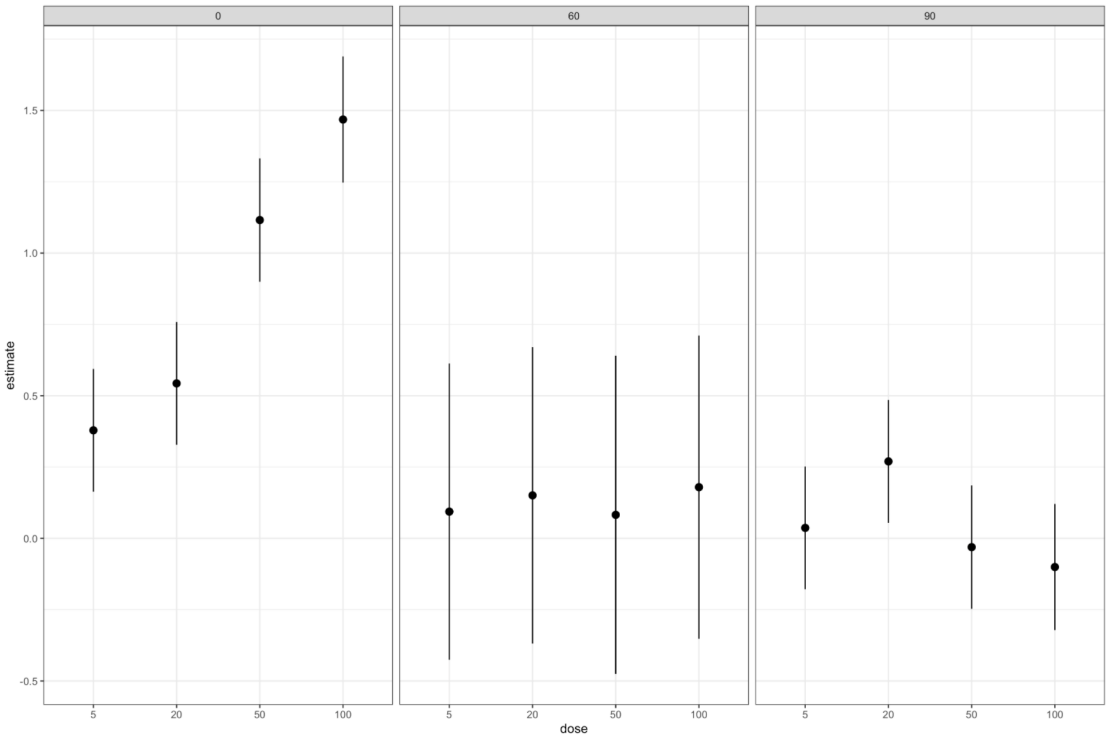

Figure 4: 95% Confidence intervals of impact of dose level (x-axis) compared to no exposure at a given repair time (panel). Each line corresponds to the 95% confidence interval of $\beta_{1,j} + \beta_{3,jk}$ at dose level j and repair time level k. Specifically, the first line in the first panel present the unit increase on log(Otmom) under dose 5 compared to no exposure if no DNA repair happens.

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	0.084	0.080	1.050	0.294	-0.073	0.241
dose5	0.379	0.110	3.454	0.001	0.164	0.594
dose20	0.543	0.110	4.948	0.000	0.328	0.759
dose50	1.116	0.110	10.116	0.000	0.899	1.332
dose100	1.468	0.113	13.027	0.000	1.247	1.689
rtime60	0.420	0.114	3.672	0.000	0.196	0.645
rtime90	0.320	0.113	2.837	0.005	0.099	0.541
sCArea	0.557	0.024	22.903	0.000	0.509	0.605
dose5:rtime60	-0.285	0.155	-1.838	0.066	-0.589	0.019
dose20:rtime60	-0.392	0.155	-2.531	0.011	-0.697	-0.088
dose50:rtime60	-1.033	0.174	-5.937	0.000	-1.375	-0.692
dose100:rtime60	-1.289	0.158	-8.145	0.000	-1.599	-0.978
dose5:rtime90	-0.342	0.155	-2.204	0.028	-0.646	-0.038
dose20:rtime90	-0.274	0.174	-1.571	0.116	-0.615	0.068
dose50:rtime90	-1.147	0.156	-7.343	0.000	-1.453	-0.840
dose100:rtime90	-1.569	0.156	-10.028	0.000	-1.876	-1.262

Figure 5: Estimation of the Final Model

	df	BIC	AdjRsquare
Otmom	17	3365.643	0.4041
Ttmom	17	3379.621	0.3866
TDNA	17	3818.005	0.3395
THDNA	17	4357.763	0.3539

Figure 6: Model comparison across log linear model with different response variables.