CENSORED DATA MODELS FOR LONGITUDINAL NEUROPSYCHOLOGICAL SCORE ANALYSIS

By

Jedediah Perkins

A THESIS

Presented to the Department of Biomedical Engineering
and the Oregon Health & Science University
School of Medicine
in partial fulfillment of
the requirements for the degree of

Master of Science

December 2011

School of Medicine

Oregon Health & Science University

CERTIFICATE OF APPROVAL

This is to certify that the Master's thesis of

Jedediah Perkins

has been approved

____

Mentor

____

Member

____

Member

TABLE OF CONTENTS

Acknowledgements

Financial support for this research was provided by Oregon Health & Science University and a National Institutes of Health TL1 training grant. Thanks go to Jeff Kaye and the Layton Aging and Alzheimer ’s Disease Center for providing access and information to longitudinal neuropsychological test data. I would like to extend special thanks to my advisor Misha Pavel for all of his help on the many projects I worked on while at Oregon Health & Science University, and to Holly Jimison for her assistance with this thesis project. Additional thanks go to Owen McCarty for his assistance in seeing that this all came together.

Abstract

Neuropsychological test scores provide a valuable tool for evaluating cognitive function and identifying cognitive decline. Unfortunately floor and ceiling effects, the methods for determining summary score information, and subject dropout (known as longitudinal censoring) are all drawbacks that inhibit the value of these tests for evaluating cognitive performance. By developing models which address these drawbacks, better estimates of cognitive performance can be obtained from score data. Earlier diagnosis, and possibly treatment, of cognitive decline may be possible with these improved score estimates.

In this thesis we utilize censored normal (Type 1 Tobit) models for longitudinal score data subject to both ceiling and floor effects. Evaluation of ceiling-afflicted data is done on the Boston Naming Test (BNT) while evaluation of floor-afflicted data is done on the Word List Delayed Recall (WLDR) test. Simulations show that failing to account for ceiling effects results in improper estimation of change points as well as population decline estimates that are significantly different than true values. Simulation shows that in longer studies failing to account for informative dropout results in an overestimation of population mean and an underestimation of population variance.

Prediction of scores at the fourth follow-up visit as well as the ability of models to classify subjects with Mild Cognitive Impairment (MCI) were evaluated for both BNT and WLDR using standard normal and Type 1 Tobit models. In the BNT, a model with quadratic decline with respect to time resulted in a classifier with an area under the Receiver Operator Characteristic (ROC) curve of 0.73 for the Tobit model, and 0.69 for standard normal, suggesting slight improvement in classification of cognitive impairment when accounting for the ceiling. Mean squared error of predicted fourth follow-up score values was 7.95 for the Tobit models and 8.05 for standard normal models. In the Word List Delayed Recall, a more widely utilized test with more data available for evaluation, a model with quadratic decline in time resulted in a classifier with an area under the ROC curve of 0.68 for Tobit models and 0.63 for standard normal models. The classifier based on the Tobit model had higher sensitivity at all ranges of specificity. Mean squared error of predicted fourth follow-up scores was 4.95 for the Tobit model and 5.35 for the standard normal model. Accounting for ceiling effects improves both classification accuracy of cognitively impaired subjects, as well as score prediction for all subjects.

1. Introduction

People are now living longer healthier lives thanks to advances in health care and major improvements in healthy lifestyles. The increases in life expectancy combined with a reduction in mortality due to other causes such as cardiovascular diseases and certain cancers have resulted in higher prevalence of neurodegenerative diseases; with cognitive decline and physical decline becoming a greater concern than ever before. Although cures for these conditions may not be readily available, early detection of cognitive decline can allow clinicians to begin treatment and mitigate the symptoms, enabling elders to live independently for much longer [1]. Because cognitive decline is progressive, it is important to be able to detect pathological declines in cognitive performance as early as possible so that treatment may begin before the decline impairs functions vital to daily independent living.

Before people develop dementia that is clearly diagnosable using current techniques, they go through a period where they begin to show signs of cognitive decline that are not sufficiently significant to interfere with daily life, although certain changes in behavior and cognitive ability may be noticed by themselves or those close to them. During this period individuals are said to have Mild Cognitive Impairment (MCI) [2]. Although intuitively appealing, the MCI condition is not well defined, and the detection of this condition depends on insensitive measurements and the subjective judgments and opinions of caregivers, clinicians and the clients. However, there is evidence that early detection of MCI allows for intervention sooner [1] and may delay the onset of symptoms of serious dementia. Currently, an individual's cognitive functionality is determined during a clinical visit, where a clinician interviews the individual and reviews his or her scores on a battery of neuropsychological and clinical tests. Even this type of assessment is triggered only after a primary care physician recommends such visits because of the complaints of the elderly individual or his or her informal caregivers. In any case these assessment visits occur at best infrequently. It would be beneficial then to be able to utilize the information from these tests over time to evaluate people's risk for developing dementia, ideally at an earlier point in time than they are currently being diagnosed with MCI. Additionally, these test scores may be used in longitudinal drug trials to evaluate the efficacy of drugs that are designed to alter the rate of change in specific cognitive functions. To that end, we should strive to develop models that accurately reflect changes in test scores over time for both healthy as well as pathological aging.

There has been a great deal of research done on estimating the rates of decline in neuropsychological and clinical measures, where a typical period between visits is six months to a year. MCI is still not well defined and the diagnostic criteria have been inconsistently applied in studies, making it difficult to compare the results [2]. There are likely different types of MCI depending on the idiopathic nature of the disease and diagnostic criteria being used. The most thoroughly studied form of MCI is amnestic MCI, which is defined by Petersen as: having a memory complaint, an objective measure of memory impairment relative to age, an otherwise preserved cognitive function, intact functional capabilities, and failing to meet the criteria for dementia [3]. Other forms of MCI include multiple-domain slight-impairment MCI and single-domain non-memory MCI. Prior researchers have attempted to address some or all of three primary challenges with longitudinal data analysis on cognitive decline: identifying from baseline examination subjects who will go on to develop cognitive impairment, estimating the underlying rate of decline for healthy versus impaired subjects, and mitigating the effects of subject dropout on estimation parameters. The diagnosis of MCI is often based off the Clinical Dementia Rating scale (CDR), a tool used by clinicians to assess the severity of dementia. The CDR score is derived from a structured interview, and has a scale ranging from 0: not impaired, 0.5: very mild dementia to 3: severe dementia.

A great deal of effort has been made to address the first challenge: identifying differences between a population that will go on to develop MCI and one that remains healthy in the near-term by utilizing baseline measurements or differences between baselines and follow-up. J. Verghese et al. used analysis of covariance on gait measures and determined that subjects who develop MCI have lower velocity and shorter stride length when walking than healthy subjects [4], while other measures such as cadence did not differentiate the groups. In this study, MCI was diagnosed by a neuropsychologist based on scoring 1.5 standard deviations below age-appropriate mean in tests of a cognitive domain, but without a diagnosis of dementia. R. Camicioli et al. used ANOVA on the “time to walk 30 feet” test. They found that subjects who develop MCI under their definition took longer to walk 30 feet than healthy subjects even at the initial visit [5], where in this case MCI was defined as a single visit with a CDR of 0.5.

Rabbitt et al. estimated practice effects as a difference between baseline and follow-up on AH4-1 intelligence test scores and found that there was a significant difference between groups stratified by intelligence on the AH4-2 test [6]. These studies have shown that are measureable differences in the score distributions between the two populations and indicate the promising possibility that test scores may be used to determine individuals who are at risk for MCI, but who have not yet reached that level of impairment.

The second challenge that has been addressed is estimating the rate of decline in test scores or cognition over time. These estimates are useful both for drug studies as well as for attempting to build better models for predicting subjects who will develop MCI. The primary goal is to estimate the rate of decline in test scores and relate it to the rate of decline of cognition, as well as be able to better predict future scores based on cognitive condition. The sequence of test scores y for an individual i at observation j can be modeled as a mixture of population and individual effects such that

$$y_{ij} = \boldsymbol{x}_{\mathrm{ij}}\boldsymbol{\beta} + \boldsymbol{z}_{ij}\boldsymbol{u}_i + \varepsilon_i \qquad (1)$$

where $\boldsymbol{\beta}$ are the coefficients for the fixed (population) effects, $\mathbf{x}_{ij}$ are the covariates for the fixed effects, $\mathbf{u}_i$ are the coefficients for the random effects, and $\mathbf{z}_{ij}$ are the covariates for the random effects, and ε_i is measurement error of the test. This is a basic linear mixed model. In longitudinal studies, the index j typically represents which clinical observation it is, starting with 0 at baseline and incrementing at each follow-up visit. The fixed effects are the non-random regression parameters that define how the population changes over time. The random effects are the parameters that determine how the individuals vary within that population, possibly with a different set of covariates, although it is common to simply use the same covariates for both the fixed and random effects. The covariates in longitudinal studies are observable measures that are likely to affect the outcome variable y, such as age, education, time since diagnosis, and comorbidity. In equation (1), variances are defined such that var(**u**) is an unknown covariance matrix, and var(ε) = $\sigma^2\mathbf{I}$, a convention which allows non-zero covariance between random effects but insures independence between fixed effects. A change point τ can be found for the population by adding it as a model parameter and fitting different curves for t<τ and t ≥τ, where t is a time measured from the clinical observation index j (common choices are age or time since baseline). Often the variable of interest is transformed so that it is linear with respect to the covariates. H. Jacqmin-Gada, D. Commenges, and J. Dartigues used a mixed model to fit scores on the Benton Visual Retention Test. Their fit was linear before an individual changepoint τ^i, and cubic after the changepoint [7]. They found that the 95% confidence interval for their prediction became wider as dementia progressed, indicating an increase in a performance variance across the population. Higher education was associated with very little decline followed by a later change point and then a very rapid decline, although it is not clear if this is simply because the more highly educated subjects were already performing at the top end of the test even after some amount of cognitive decline, resulting in a test design that masked the decline in the highly educated subjects. D. Howieson et al. defined MCI as 2 consecutive visits with a CDR of

0.5. They examined four measures: Wechsler Memory Scale, Logical Memory I and II Story A, category fluency for animals, and Block Design. For each measure they fit a separate longitudinal mixed model with two linear declines separated by a population change point τ for MCI subjects, and a single linear model for healthy subjects. All four measures indicated that τ occurred prior to the date of diagnosis [8]. Logical memory tests showed differences between the groups prior to the change point, with healthy groups actually improving on the tests as they aged. This could be due to a practice effect, which is common amongst healthy individuals. P. Rabbitt et al. determined that the degree of such practice effects in healthy individuals depends on age and baseline score ability [6]. They also noted both floor and ceiling effects that resulted in nonlinearities in the practice effects. This was because subjects who performed very poorly were the ones who were already much declined, and showed very little practice effect due to that, but subjects who performed very well couldn't really improve because they were already near the top and so they didn't show much improvement either. Subjects whose performance was in the middle showed the greatest increase in score on follow-up due to practice effects. A. Zehnder et al. noted that subjects who went on to develop dementia in the BASEL cohort did not show practice effects while those that remained healthy did [9] but that the effect did not improve classification over just using baseline score data.

The third challenge that researchers attempt to address is missing data and subject dropout. Missing measurements and subject dropout are common forms of censoring in longitudinal studies. If the missing data is not Missing

Completely at Random (MCAR), then not modeling the missing observations leads to biased estimates of the population parameters. MCAR the missing observations are independent of both observable variables and unobservable parameters. Other forms of missing data are particularly problematic when trying to identify a disease population, especially if the probability of dropout is dependent on variables, observable or not, relating to the disease. The measurements can be considered Missing at Random (MAR) when the distribution of missing data d_{mi}, depends only on observed outcomes y_i^{obs}, or it can be Missing Not at Random (MNAR) where the distribution of missing data can depend on both observed and unobserved outcomes. Data that is MAR is managed by modeling the dependencies and testing on the observations.

Several approaches have been taken to model missing data, and if it is handled poorly (such as Last Observation Carried Forward, a method where all missing observations after dropout are assumed to have the value of the last non-missing observation) it can lead to incorrect conclusions [10], [11]. Dropout can leave the remaining data with a non-normal distribution, in which case a median regression model may provide better estimators [12]. L. Su and J. Hogan developed varying-coefficient models (VCMs) that differentiated between administrative and other dropout methods [13]. Yuan and Little [14] implemented Mixed-Effect Hybrid Models (MEHMs) to jointly model the mechanism for missing data and the outcome process. Their model is based on a specific factorization of $f(D_i,\mathbf{y}_i,\mathbf{b}_i|\mathbf{x}_i,\boldsymbol{\theta})$, the joint distribution of the observed outcomes $\mathbf{y}_i$, the random effects $\mathbf{b}_i$, and the dropout times D_i on covariates $\mathbf{x}_i$ and model parameters $\boldsymbol{\theta}$, where i is a subject index. In a simulated data study they found that the MEHM provided nonbiased estimates of population parameters when the data was MAR and MNAR, but normally distributed. Unfortunately, in data with ceiling and floor effects, such as the scores of many neuropsychological tests, the normality assumption often does not hold true and recorded scores may in fact be subject to censoring whereby the test is too easy for a subset of the population, and they score at the ceiling. D. Hedeker and R. Gibbons applied pattern-mixture models with missing data patterns as a group effect to address longitudinal censoring in Inpatient Multidimensional Psychiatric Scale Item 79 (IMPS79) data [15], with improvements in parameter estimates over models that did not include missing data patterns.

A great deal of work has gone in to addressing the issues of identifying differences in the test scores between healthy and cognitively impaired subjects, as well as examining the rate of decline of those scores as a function of healthy or pathological gaining under longitudinal censoring. Unfortunately, these methods assume that the scores are a direct representation of the cognitive function of interest without imposing any within-test censoring. This is not the case for tests that contain a fixed number of items, such that the score has a maximum possible value of C (the ceiling) and a minimum possible value of F (the floor). An example of a test with a significant ceiling effect is in Figure 1. Failing to account for the ceilings and floors in these tests can lead to incorrect estimates of changepoints and rates of change. Addressing the issue of ceiling and floor effects on longitudinal data is the focus of this thesis.

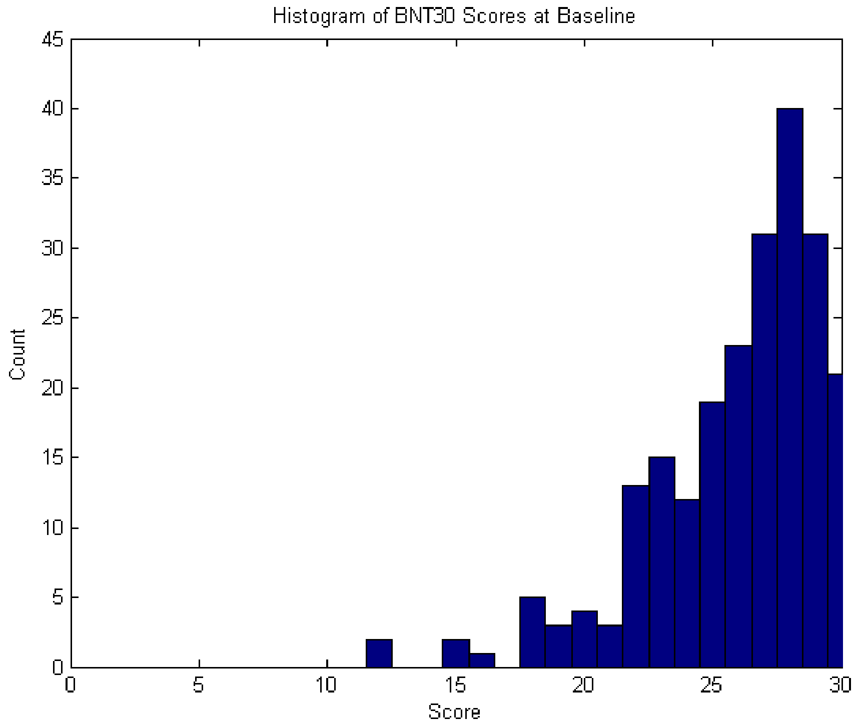

Figure 1 – Example of a neuropsychological test with significant ceiling effects: the 30-item Boston Naming Test, where a score of 30 is the highest possible, and many subjects achieve it.

2. Hypotheses and Research Aims

Tests with clear ceiling and floor effects are those that are made up of a set number of items, resulting in a range of possible integer scores. Due to this scoring system, they appear to be binomially distributed, although assuming that they can then be modeled as normally distributed due to the normal approximation of the binomial distribution is a potential pitfall. There has been a limited amount of prior exploration of within-test censoring due to floor and ceiling effects. H. Dodge et al. utilized a pattern-mixture model to address missing data bias in the Consortium to Establish a Registry for Alzheimer's Disease (CERAD) Word List Delayed Recall (WLDR) over a 12-year longitudinal study. In their model they accounted for the floor and ceiling effects of the WLDR, but did not explore if accounting for them had any effect on their outcomes [16]. More recently, L. Wang et al. showed that data that changed at a constant linear rate with an imposed ceiling would appear nonlinear if the ceiling was not accounted for [17]. In their work, they were concerned with short term test growth over repeated measurements with fixed time intervals. They showed that not accounting for the ceiling on the test lead to an underestimation of the population mean and the fixed rate change parameters of the growth curve model. B. Uttl explored the effects of ceilings on test validity and reliability for tests with a presumed normal distribution and significant ceiling proportion [18]. He found that intertrial correlation was smaller for shorter tests with larger ceiling proportion, and that correlations between shorter and longer tests designed to measure the same thing were improved if the ceiling effects were accounted for.

In this work, I will address the results of not accounting for floor and ceiling effects on longitudinal decline analysis with variable test-retest intervals. When not accounting for the floor and ceiling effects, I assume that the outcome variable would then instead be assumed normally distributed. Otherwise, I utilize a censored normal model [19], where the true distribution of the variable of interest if there were no floor or ceiling on the test is

$$y_{ij}^{*} = \boldsymbol{x}_{\mathrm{ij}}\boldsymbol{\beta} + \boldsymbol{z}_{ij}\boldsymbol{u}_i + \varepsilon_i \tag{2}$$

but the scores are censored at the floor F and the ceiling C such that the scores that are actually observed are

$$y_{ij} = \begin{cases} C & y_{ij}^{*} \geq C \\ y_{ij}^{*} & F < y_{ij}^{*} < C \\ F & y_{ij}^{*} \leq F \end{cases} \tag{3}$$

This is known as a Type I Tobit model. When fitting the changes of the underlying model in (2) over time subject to the ceiling and floor effects present in (3), I will use the term “Tobit Decline Model”. There are three hypotheses being tested in this work:

Hypothesis 1: In ceiling-censored normally distributed data, parameters of single timepoint regressions are miss-estimated when the ceiling is not accounted for.

Hypothesis 2: In ceiling-censored normally distributed data with a longitudinally declining population, the rate of population decline is underestimated when the ceiling is not accounted for.

Hypothesis 3: When fitting a longitudinal decline model to two populations where the observed measurement is censored at the floor and ceiling, not accounting for the censoring at floor and ceiling leads to greater errors in score prediction and a reduction in classification accuracy between the two populations.

The first aim of this thesis is to show through simulation that the Tobit Decline Model can be utilized to more accurately estimate single-timepoint parameters, rate of decline, and decline changepoint than a standard linear regression model.

The second aim of this thesis is to show that the Tobit Decline Model estimates the observed scores in empirical test data better than a standard linear mixed model, as well as improves classification accuracy of Mild Cognitive Impairment (MCI). Two tests will be examined. The first test is the 30-item Boston Naming Test (BNT30) which has a significant ceiling proportion at baseline. The second test is the WLDR test, which has subjects who perform at both the floor and ceiling.

3. Fitting and Model Selection Methods

As the primary concern of this thesis is the effect of not accounting for ceiling and floor effects in a model and not the methods used to fit that model, I will not explore the different methods for fitting longitudinal Tobit Decline Models. Since in future work it is likely I will utilize distributions that do not have closed form analytical solutions, the model fitting in this paper will be done with Markov Chain Monte Carlo (MCMC) methods. Model parameters are estimated using Bayesian inference Under Gibbs Sampling with the WinBUGS software [20]. In Gibbs sampling, it is necessary to provide prior distributions for all of the model parameters, although those priors need not be conjugate or informative. An initial estimate of the model parameters $\boldsymbol{\theta}^{(0)}$ is chosen, and n_k samples are generated by sampling $\theta^{(k)}$ from

$$p(\theta_l^{(k)}|\theta_1^k, \dots, \theta_{l-1}^k, \theta_{l+1}^{k-1}, \dots, \theta_{nl}^{k-1}, y) \qquad (4)$$

for the n_l parameters. The model parameters $\boldsymbol{\theta}$ in the Tobit Decline Model as well as standard normal mixed models are the fixed and random effect coefficients $\boldsymbol{\beta}$ and **u** from equation (1). After a burn-in period where parameter estimates are still converging towards the true values, the samples of $\boldsymbol{\theta}$ converge to a stable estimate and for large k the samples approximate the joint distribution $p(\boldsymbol{\theta}|y)$. By Bayes' theorem:

$$p(\boldsymbol{\theta}|y) \propto p(y|\boldsymbol{\theta})p(\boldsymbol{\theta}) \qquad (5)$$

where $p(\boldsymbol{\theta})$ is the prior probability on the model parameters and $p(y|\boldsymbol{\theta})$ is the likelihood. In this work I use non-informative conjugate priors for the model parameters.

In this thesis, I will be comparing the performance of censored normal models with standard normal models. For the standard normal mixed model described in (1), the likelihood is

$$p(y_{ij}|\boldsymbol{\beta},\boldsymbol{u}_i) = \prod_{i,j} \varphi\left(\frac{y_{ij} - (\boldsymbol{x}_{ij}\boldsymbol{\beta} + \boldsymbol{z}_{ij}\boldsymbol{u}_i)}{\sigma}\right)/\sigma \tag{6}$$

For the Tobit Decline Model resulting from (2) and (3), the likelihood is

$$\begin{gathered} p(y_{ij}|\boldsymbol{\beta},\boldsymbol{u}_i) = \prod_{y_{ij}=F} \Phi\left(\frac{F - (\boldsymbol{x}_{ij}\boldsymbol{\beta} + \boldsymbol{z}_{ij}\boldsymbol{u}_i)}{\sigma}\right) * \prod_{y_{ij}=C} \Phi\left(\frac{(\boldsymbol{x}_{ij}\boldsymbol{\beta} + \boldsymbol{z}_{ij}\boldsymbol{u}_i) - C}{\sigma}\right) * \\ \prod_{y_{ij}\neq C,F} \varphi\left(\frac{y_{ij} - (\boldsymbol{x}_{ij}\boldsymbol{\beta} + \boldsymbol{z}_{ij}\boldsymbol{u}_i)}{\sigma}\right)/\sigma \end{gathered} \tag{7}$$

where Φ() is the standard normal cumulative distribution function and φ() is the standard normal probability distribution function and $\mathbf{u}_{ij}$ is distributed multivariate normal MVN(0,**Σ**).

When determining which model is a better fit, it is important to consider not only the likelihood but the number of effective parameters in the model, as increasing the number of parameters will tend to improve likelihood even if it may just be an artifact of over fitting. In order to determine which of the two models better fit the data, deviance information criterion (DIC) is used. In DIC, the deviance of a model is

$$D(\boldsymbol{\theta}) = -2\log(p(y|\boldsymbol{\theta})) \tag{8}$$

and the effective number of parameters is

$$p_D = E^{\theta}[D(\boldsymbol{\theta})] - D(\overline{\boldsymbol{\theta}}) \tag{9}$$

Where E[] is the expected value. The DIC is

$$DIC = p_D + E^{\theta}[D(\boldsymbol{\theta})] \tag{10}$$

The larger the DIC, the worse the model fit is. The model with the lowest DIC is the one that is considered the correct model for the given data, although small differences in DIC are not conclusive due to the value being directly calculated from the samples of the Monte Carlo simulation such that a small shift in one sample results in a small shift in DIC.

4. Effect of Censoring on Parameter Estimation

Not accounting for ceilings in normally distributed data can lead to a variety of potential problems when attempting to fit a model to the data. In this chapter I will address three ways in which parameters are incorrectly estimated when fitting a model assuming normally distributed data and not accounting for the ceiling: miss-estimation of parameters in single timepoint analysis, underestimation of the rate of decline in longitudinal data, and improper estimation of changepoints in longitudinal decline data. Simulated data was generated in MATLAB and parameter estimation was performed in WinBUGS.

4.1 Single Timepoint Parameter Estimation Subject to Ceilings

4.1.1 Methods

In order to explore the effect of ceilings on single timepoint parameter estimation, I simulated a sample population and varied the proportion of the observables that were cut off by the ceiling. A population of N=300 was generated such that

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon \tag{11}$$

where $\beta_0 = 52$, $\beta_1 = -1$, $\beta_2 = 2$, x_1~N(20,25), x_2~N(10,4), and ε~N(0,1). This resulted in y being distributed N~(52,42) prior to the application of the ceiling. The ceiling was varied so that the proportion of data at the ceiling ranged from 6.7% to 33% over 5 separate trials.

The models fit in WinBUGS were a simple linear regression of the same form in (11), as well as a censored normal (Type 1 Tobit) regression model where the ceiling level was known in the censored regression model. The prior distributions of the parameters were β_0, β_1, β_2~N(0,1E-6) and $1/\sigma^2$~Γ(1,1). 100 repetitions were run at each trial. 1000 samples were used for burn-in during the Gibbs sampling, and then the next 1000 samples were taken as the parameter estimates.

4.1.2 Results

Over all trials the censored normal model had a lower deviance information criterion (DIC) than the linear regression model. Average DIC values as well as the means and standard deviations of the parameter estimations under each model can be found in Table 1. Figure 2 shows the estimates of each parameter for each model as a function of proportion of data at the ceiling.

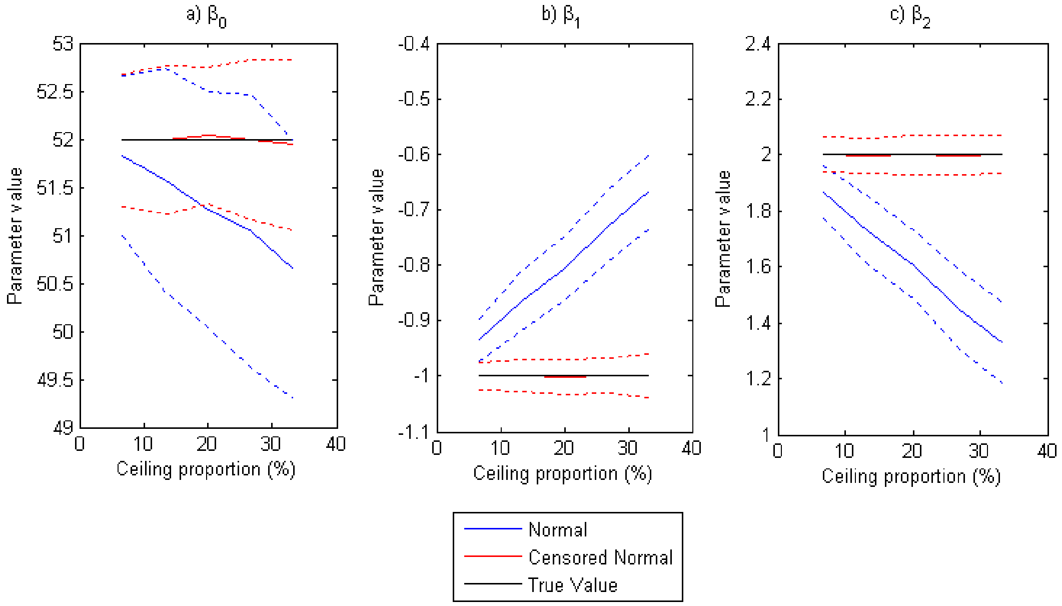

Figure 2 – Parameter estimates and their 95% confidence intervals for both the censored normal model and the standard normal model as a function of ceiling proportion.

The censored regression model estimates the parameters very close to their true values even with as much as 33% ceiling proportion (Table 1), with all 95% confidence intervals for the estimates containing the true values. The standard normal regression model is very sensitive to ceiling effects and the true values β_1 of β_2 and fall outside the 95% confidence intervals of the estimates. The 95% confidence interval of the estimate of β_0 contains the true value out to 33%, but the trend indicates that it will fall outside at any higher ceiling proportion.

	True	Normal					Censored Normal				
Ceiling Proportion (%)	NA	6.7	13.3	20.0	26.7	33.3	6.7	13.3	20.0	26.7	33.3
DIC	NA	990	1100	1169	1225	1254	806	757	705	654	599
β_0	52	51.8 (.42)	51.6 (.59)	51.3 (.63)	51.0 (.72)	50.6 (.69)	52.0 (.35)	52.0 (.39)	52.0 (.37)	52.0 (.43)	52.0 (.45)
β_1	-1	-.93 (.02)	-.86 (.03)	-.80 (.03)	-.73 (.03)	-.67 (.03)	-1.0 (.01)	-1.0 (.01)	-1.0 (.02)	-1.0 (.02)	-1.0 (.02)
β_2	2	1.9 (.05)	1.7 (.06)	1.6 (.06)	1.4 (.07)	1.3 (.07)	2.0 (.03)	2.0 (.03)	2.0 (.04)	2.0 (.04)	2.0 (.04)
σ	1	1.3 (.11)	1.5 (.12)	1.7 (.12)	1.9 (.12)	1.9 (.11)	1.0 (.05)	1.0 (.05)	1.0 (.04)	1.0 (.05)	1.0 (.05)
μ_y	52	51.8 (.08)	51.6 (.1)	51.3 (.12)	50.9 (.16)	50.6 (.18)	52.0 (.06)	52.0 (.06)	52.0 (.06)	52.0 (.08)	52.0 (.08)
Var(y)	42	37.4 (1.2)	32.9 (1.5)	29.4 (1.5)	25.3 (1.5)	22.0 (1.6)	42.1 (.9)	41.9 (1.0)	42.1 (1.2)	41.9 (1.1)	42.0 (1.3)

Table 1 – Comparison of true population parameters with estimates from normal and censored normal models. Numbers in () are the standard deviations of the estimates.

The model fits generate an estimate for the pre-censoring distribution of the observable y as well. The estimates of the mean and variance of y for all trials are found in Table 1.

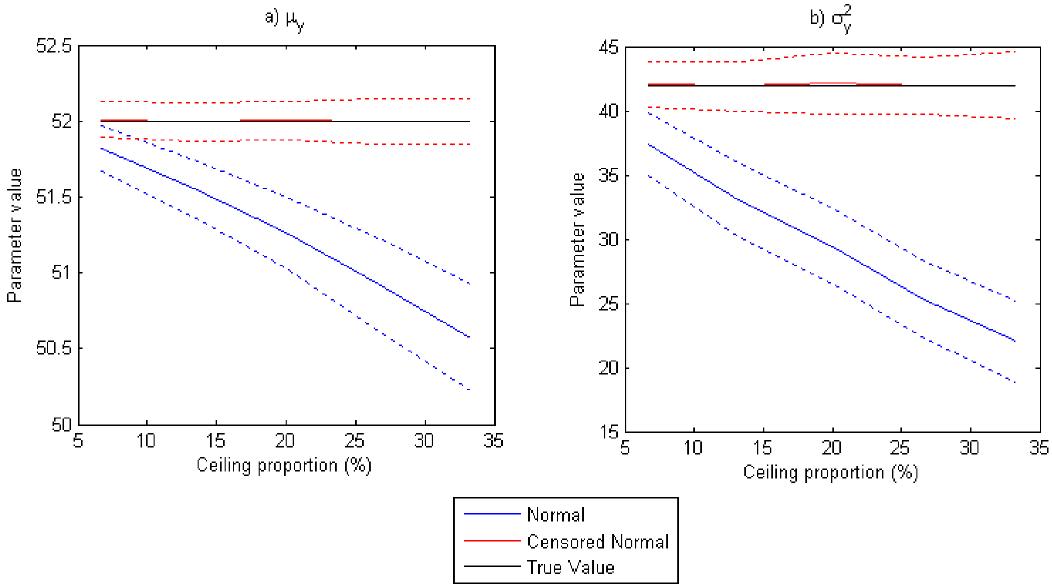

Figure 3 – Estimates of the mean and variance of the observable y as well as their true values as a function of ceiling proportion.

Figure 3 shows the estimates of the mean and variance as a function of ceiling proportion. The censored normal model closely represents the true distribution while the standard normal model underestimates both the mean and variance at all ceiling proportions tested. The 95% confidence intervals for both estimates of the standard normal model do not contain the true values.

4.2 Rate of Longitudinal Decline Subject to Ceilings

4.2.1 Methods

Next, to explore the effect of ceilings on estimation of the rate of decline, I simulated a normally distributed sample population with a fixed proportion (15%) of the observables that were cut off by the ceiling at baseline and declined the population at a constant rate across all individuals for a period of four years which covered five total samples for each simulated subject. A population of N=300 was generated such that

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2i} + \varepsilon \tag{12}$$

where β_0 = 52, β_1 = -3, β_2 = 2, x_{1i0}~N(0,16), x_2~N(10,4), and ε~N(0,1). This resulted in y being distributed N~(72,151) at baseline, prior to the application of the ceiling. Four additional time samples were generated for each individual i such that $x_{1ij} = x_{1i0}+0.8j$.

The models fit in WinBUGS were the same form of linear regression from the section 4.1. The prior distributions of the parameters were once again β_0, β_1, β_2~N(0,1E-6) and $1/\sigma^2$~Γ(1,1). 100 repetitions were run, with 1000 samples used for burn-in and the next 1000 samples were taken as the parameter estimates.

4.2.2 Results

The censored normal model (DIC = 3862) fit the data better than the standard normal model (DIC = 6719). Parameter estimates for each of the models are shown in Table 2. Once again, all of the true parameters of the distribution were within the 95% confidence intervals of the estimates of the censored normal model. For the standard normal

	True	Normal	Censored Normal
β_0	52	52.9 (0.6)	52.0 (0.1)
β_1	-3	-2.7 (.06)	-3.0 (.01)
β_2	2	1.8 (.07)	2.0 (.01)
σ	1	2.3 (.3)	1.0 (.02)

Table 2 – Comparison of true population parameters with estimates from normal and censored normal models. Numbers in () are the standard deviations of the estimates.

model, the only parameter that contained the true value within the 95% confidence interval was β_0.

Figure 4 shows the estimates of the mean of the population as a function of time. The censored normal model captures the mean within its 95% confidence interval over the entire time period.

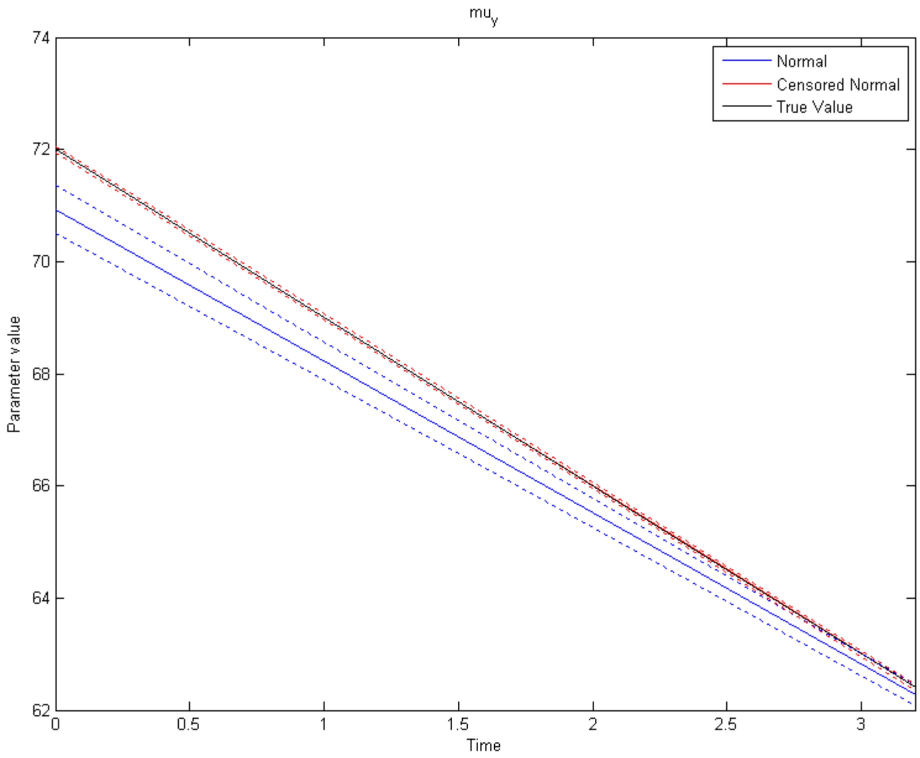

Figure 4 – Estimate of mean score as a function of time.

The mean is outside the 95% confidence interval predicted by the standard normal model, but as time increases the estimate trends closer to the true value. This is due to more of the data dropping below the ceiling threshold.

4.3 Changepoint Estimation Subject to Ceilings

4.3.1 Methods

To explore the effect of ceilings on estimation of a changepoint, I simulated two sample populations with the same ceiling, but the population parameters at baseline were different to generate different ceiling proportions of the observables as most changepoint analyses are looking at two populations under the same testing conditions and trying to determine when a changepoint occurs in each. Two populations of N=300 were generated such that

$$y_{ij} = \begin{matrix} \beta_0 + \beta_1 t_j + \varepsilon & t_j < \tau \\ \beta_0 + \beta_1 \tau + \beta_2 \left(t_j - \tau\right) + \varepsilon & t_j \geq \tau \end{matrix} \tag{13}$$

where $\beta_0 = 56$ for population 1 and 59 for population 2, $\beta_1 = -1$, $\beta_2 = 3$, $\tau = 4.5$ and $\varepsilon \sim N(0,20)$. The ceiling was set at 60, which resulted in ceiling proportions of 20% and 40%. Later time samples were generated at $t_j = 2j$ for j=1…6.

The models fit in WinBUGS were piecewise linear separated by a random population changepoint. The prior distributions of the parameters were τ, β_0, β_1, $\beta_2 \sim N(0,1E\text{-}6)$ and $1/\sigma^2 \sim \Gamma(1,1)$. 100 repetitions were run, with 1000 samples used for burn-in and the next 1000 samples were taken as the parameter estimates.

4.3.2 Results

Under both initial means, the censored normal model had a lower DIC than the normal linear regression. Average DIC values as well as the means and standard deviations of the parameter estimations under each model can be found in Table 3.

	True (Pop. 1)	Normal (Pop. 1)	Censored Normal (Pop. 1)	True (Pop. 2)	Normal (Pop. 2)	Censored Normal (Pop. 2)
DIC	NA	11341	10950	NA	11256	10375
β_0	56	55.6 (1.2)	56.4 (1.1)	59	57.1 (1.4)	58.9 (1.1)
β_1	-1	-1.1 (.03)	-1.2 (.03)	-1	-.82 (.04)	-1.1 (.05)
β_2	-3	-3.0 (.02)	-3.0 (.01)	-3	-3.1 (.06)	-3.1 (.06)
σ	4.5	5.6 (.75)	6.0 (.68)	4.5	5.5 (.83)	6.3 (.70)
τ	4.5	4.9 (.24)	4.7 (.18)	4.5	4.7 (0.20)	4.5 (0.16)

Table 3 – Comparison of true population parameters with estimates from normal and censored normal models. Numbers in () are the standard deviations of the estimates.

As is apparent in Table 3, neither model was able to accurately estimate the standard deviation of the error term. However, the censored normal model's estimate of the changepoint for both trials contained the true value within the 95% confidence interval. The normal model underestimated the population mean in the second (40% of observables censored) trial.

5. Application to Test with Significant Ceiling Proportion: Boston Naming Test

Since it is clear from simulations that both baseline population and longitudinal decline parameters are poorly estimated by normal linear regression models, it is valuable to compare the performance of censored normal Tobit Decline models with standard normal linear mixed models on empirical data with significant ceiling effects. The test that I will be using that has a significant ceiling effect is the 30-item version of the Boston Naming Test (BNT30). The BNT30 is a confrontation naming test where subjects are presented with line drawings and must identify what they represent. The score is the sum of correct answers. There is a significant ceiling effect on this test due to high probability of correctly identifying the items on the test.

The data for this section is pooled from four long term longitudinal studies conducted by the Layton Aging and Alzheimer 's Center at Oregon Health & Science University. The studies are the Oregon Brain Aging Study (OBAS), the Klamath Exceptional Aging Project (KEAP), the African American Dementia and Aging Project (AADAPt), and the Intelligent Systems for Assessment of Aging Changes study (ISAAC). OBAS began in 1989, with an initial recruitment of very healthy adults aged 55 or older. A second arm was added in 2004, with subjects aged 85 or older who were of more average health than the initial arm. Subjects are seen annually and administered a battery of neuropsychological tests and given a full clinical evaluation. AADAPt follows African American individuals aged 65 or older and administers a battery of neuropsychological tests every 6 months. KEAP began in 1999 as an effort to add a rural population to those already being studied at the Layton Aging and Alzheimer's Center. ISAAC is a 5-year study that has enrolled 164 subjects ages 70 and up in the Portland, OR metro area. Subjects received a computer and free internet upon enrollment in the study and undergo continuous in-home monitoring with motion sensors and computer games. Additionally, they receive a battery of neuropsychological tests annually as well as a full clinical diagnosis. A large, diverse population can be examined by pooling the data across these studies.

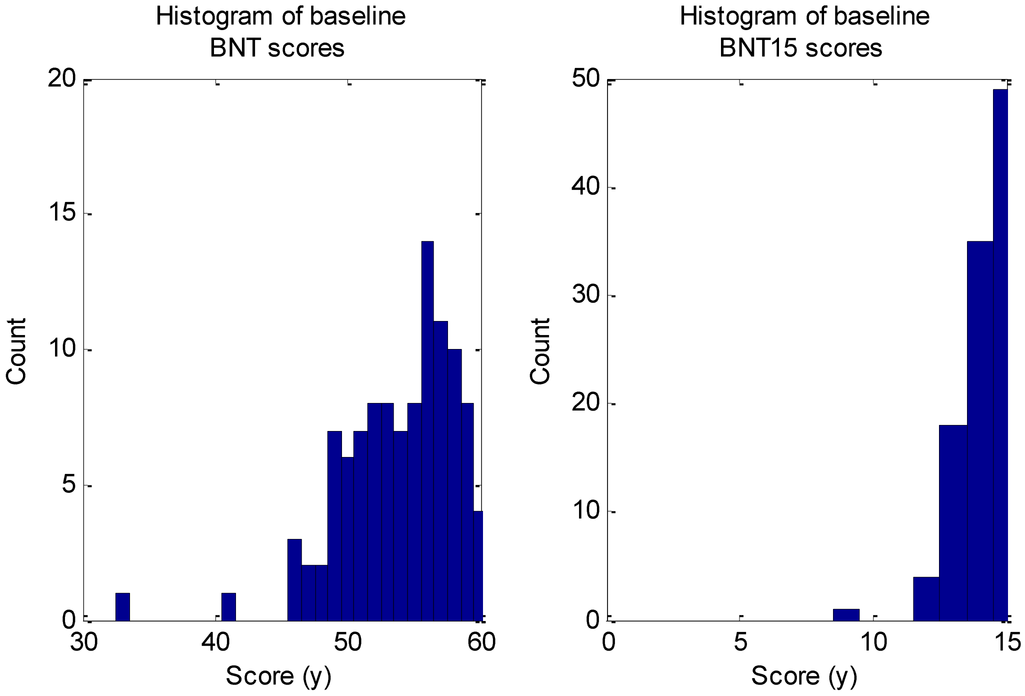

Figure 5 – Comparison of BNT60 and BNT15 scores for the same subjects at baseline.

It was a preliminary analysis of the 15 and 60-item BNTs that indicated a promise for Tobit Decline Models to improve estimation. The BNT60 shows less of a ceiling effect than the 15 or 30-item versions, and despite reports that the psychometric properties are the same on the shorter versions of the test [21], Figure 5 shows a significantly different distribution for the 15 and 60 item versions. The Pearson's correlation coefficient of just the scores at baseline is r=0.68. However, when both sets of baseline scores were fit to a censored normal distribution, the Pearson's correlation coefficient between the latent scores was r=0.98 and the shapes of the distributions were more similar (Figure 6). This seemed to indicate that censoring in the shorter BNT15 was cutting off performance at a level below what was measured by the longer BNT60.

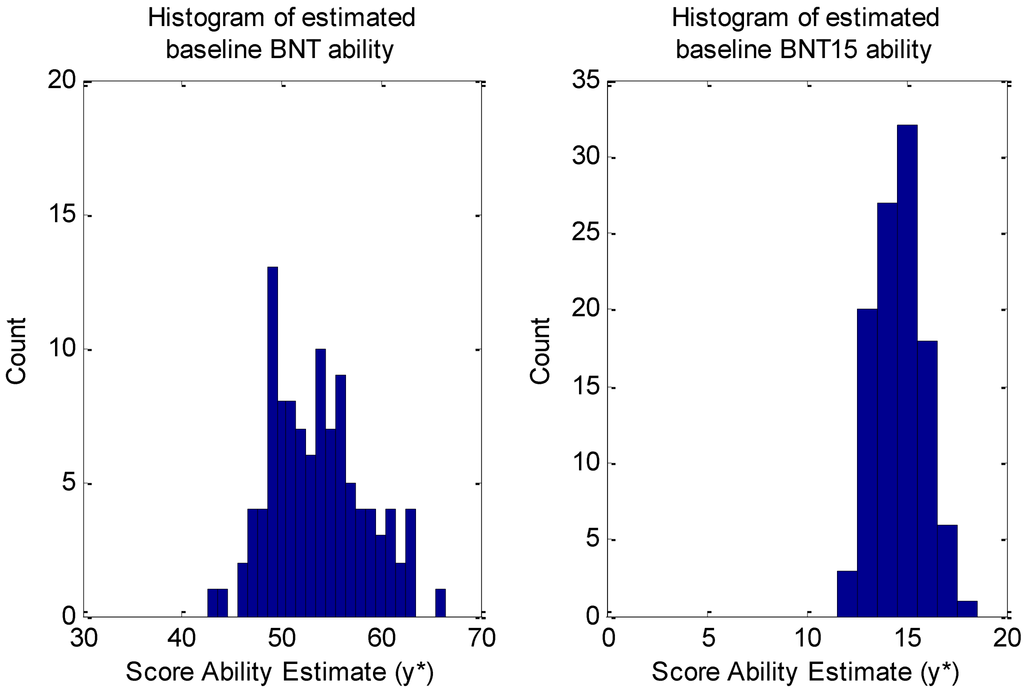

Figure 6 – Comparison of baseline distribution of the latent score y* when fit to a censored normal distribution.

For this work, I am only considering subjects who were not cognitively impaired during enrollment in the study and who have at least three follow-up visits in which a BNT30 score was recorded. Three follow-ups is a short time period, but it was necessary to maintain the statistical power in the analysis.

Furthermore, for the prediction of mild cognitive impairment (MCI) it is unlikely that a subject will have many visits where they take a full neuropsychological battery before they develop cognitive impairment, given how low the frequency the administration of those test batteries is. MCI will be defined as two consecutive visits with a CDR of 0.5 or greater. Across all of the studies, N=225 subjects meet the criteria, N_{MCI}=25 of which have developed MCI. Figure 7 shows a histogram of baseline BNT30 scores from the subjects who meet these requirements; the ceiling effect is quite evident.

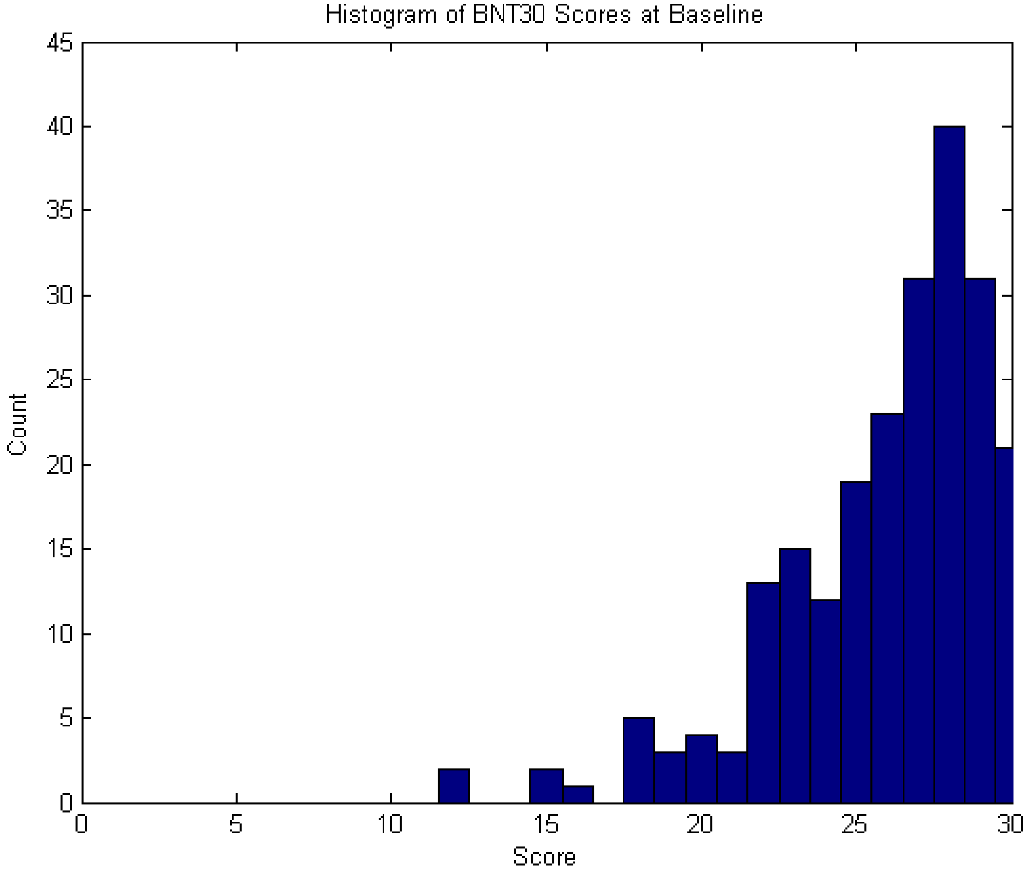

Figure 7 – Histogram of the baseline BNT30 scores for the subjects meeting the criteria for analysis.

5.1 Decline Trajectories with Significant Ceiling

5.1.2 Methods

Estimating the differences in rates of decline between healthy and pathological aging can provide a means for identifying cognitively impaired subjects as well as validate the effects of treatments. Here I will compare two models of decline, one where both healthy and cognitively impaired subjects decline linearly with respect to time, but at different rates, and another where the decline of cognitively impaired subjects is quadratic with respect to time. The mixed model employed is:

$$y_{ij}^{*} = (u_{0i} + \beta_0) + (u_{1i} + \beta_1)x_{1i} + (u_{2i} + \beta_2)x_{2ij} + MCI(u_{3i} + \beta_3)x_{3ij} + \varepsilon_i \qquad (14)$$

where MCI is a group indicator that is 1 for MCI subjects and 0 for healthy subjects, x_{1i} the age at baseline, x_{2ij} is the time elapsed at visit j since baseline, and x_{3ij} is either the time elapsed or the square of the time elapsed, depending on the model. In this model the ceiling C=30 and the floor F=0. Initially education was included as a possible covariate as well but the coefficient was 0 so it was dropped from the model. The subject-specific coefficients $\mathbf{u}_i$ are assumed to be distributed mean 0 with covariance matrix **Σ**, and ε is assumed to be distributed N~(0,σ^2).

Fitting the models in WinBUGS, the prior distributions of the parameters were β_0, β_1, $\beta_{2,}$ β_3~N(0,1E-6) and $1/\sigma^2$~Γ(1,1). The prior distribution of **Σ** is ~Wishart(**I**,4). For each model 1000 samples used for burn-in and the next 1000 samples were taken as the parameter estimates.

5.1.2 Results

The results of the fits for the quadratic and the linear models are listed in Table 4. If the 95% confidence interval for the parameter estimate includes 0, the parameter is negligible and dropped from the model. The results in Table 4 indicate that there is post-baseline time related decline for healthy subjects as a group. This could be due to the short time window of the analysis (average time from baseline to last follow-up 3.9 years). The random-effects of the model indicate an increased variance associated with age in all subjects. All decline models showed a time dependent decline for the MCI group. When adjusting for the parameters that drop out, model reduces to

$$y^*_{ij} = \beta_0 + \beta_1 x_{1i} + u_{2i} x_{2ij} + MCI\beta_3 x_{3ij} + \varepsilon_i \qquad (15)$$

The implications of this model are that for subjects who are not at risk for developing MCI the test is simple enough that they do not show any age-related decline, but subjects who are at risk for MCI have an annual decline that is estimated to be slightly faster under the Tobit Decline Model on average.

	Linear (Tobit)				Linear (Normal)				Quadratic (Tobit)				Quadratic (Normal)			
DIC	3628				3732				3632				3735			
β$_0$	31.5(1.9)				33.0(2.5)				28.2(1.9)				33.0(2.3)			
β$_1$	-.06(.02)				-.08(.03)				0				-.10(.03)			
β$_2$	0				0				0				0			
β$_3$	-1.1(.22)				-1.0(.21)				-.33(.09)				-.31(.09)			
σ	1.7(.06)				1.6(.05)				1.7(.06)				1.6(.05)			
Σ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	.55 (.11)	0	0	0	.44 (.09)	0	0	0	.25 (.09)	0	0	0	.26 (.09)	0
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table 4 – Parameters estimates of the linear and quadratic mixed models on the 225 subjects with BNT30 data. Values in the () are the standard deviations of the estimates.

5.2 Score Prediction and Identification of Mild Cognitive Impairment

5.2.1 Methods

The two primary purposes for developing cognitive decline models in aging populations are to identify subjects who will develop cognitive impairment, and to improve score prediction for model validation in treatment efficacy studies. With that in mind, both classification and score prediction can be analyzed together. I begin by classifying subjects using a Naïve Bayes classifier and computing the log of the likelihood ratio

$$ln\left(\frac{p(MCI|y_i)}{p(h|y_i)}\right) = ln\left(\frac{p(MCI)}{p(h)}\right) + \sum_j ln\left(\frac{p(y_{ij}|\theta_{MCI})}{p(y_{ij}|\theta_h)}\right) \qquad (16)$$

where p(MCI) and p(h) are the prior probabilities of MCI and healthy populations, respectively. Then, $p(y_{ij}|\theta_{MCI})$ is the likelihood (equation 6) of the model when MCI=1, and $p(y_{ij}|\theta_h)$ is the likelihood of the model when MCI=0. The classifier is trained on a training set so that the probability distributions $p(MCI|y_i)$ and $p(h|y_i)$ can be estimated. Subjects in the evaluation set are then classified using the ratio of these estimated distributions. They are assigned to the MCI group if the log-likelihood ratio is greater than a threshold, and are considered healthy otherwise. The priors are estimated from the proportions of MCI and healthy subjects in the training set, and the classification threshold is varied to generate a Receiver Operating Characteristic (ROC) curve. The ROC curve provides a visual representation of the sensitivity of the classifier as a function of reduction in specificity. The area under an ROC curve is a measure of the quality of a classifier. A perfect classifier always correctly assigns groups and has an area under the ROC curve of 1. A completely random guess would result in a linear ROC curve with slope 1 and an area under the curve of 0.5. For this analysis, the training set consisted of the subjects who only completed four score evaluations (N_{tr}=278, N_{trMCI}=13). The evaluation set consisted of subjects who had completed at least five score evaluations, although it was only evaluated on their fifth (N_{ev}=47, N_{evMCI}=12).

To evaluate score prediction, the evaluation set's subject scores are then predicted for their fifth evaluation which occurs at time t_{i5} after baseline and the predicted score is compared to their actual score. The predicted score value depends on the parameters of the model, which is in turn dependent on the classification. The score predictions are evaluated by determining the mean squared error of the score fifth evaluation score predictions under the prior probability distribution resulting in 80% sensitivity to MCI.

5.2.2 Results

ROC curves for the classification under the linear and quadratic models from section 5.1 are shown in Figures 8 and 9, respectively. The areas under the ROC curve (specified AUC) as well as the specificity of the model when the priors are chosen for 80% sensitivity to prediction of MCI as a positive are reported in Table 5. Mean squared error of the score fifth evaluation score predictions under the prior probability distribution resulting in 80% sensitivity to MCI are reported in Table 5.

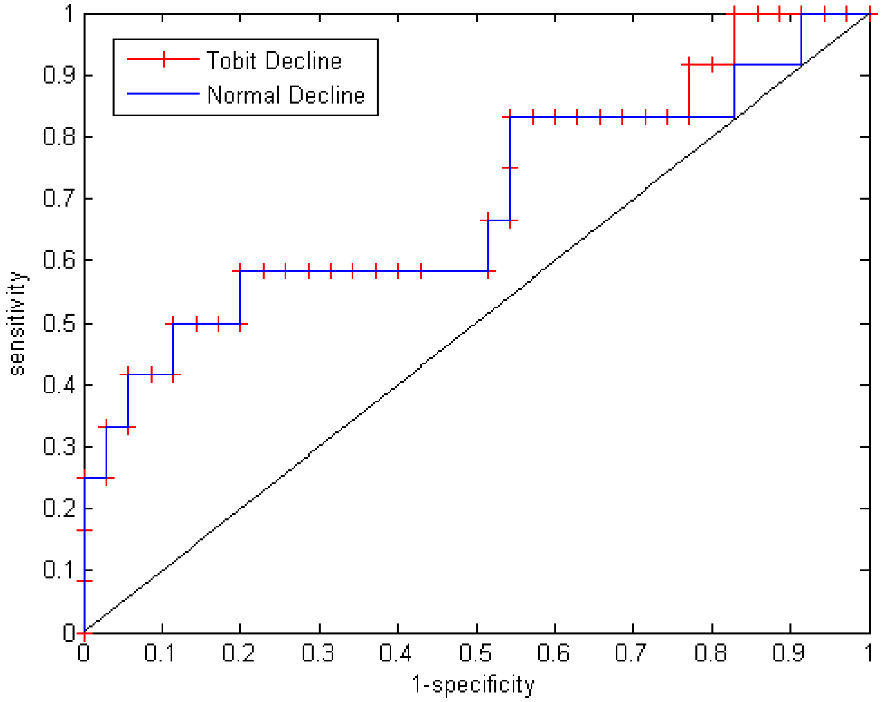

Figure 8 – ROC curve of the linear-in-time decline models for the Tobit Decline Model and the normal decline model. The black line is a totally random classifier.

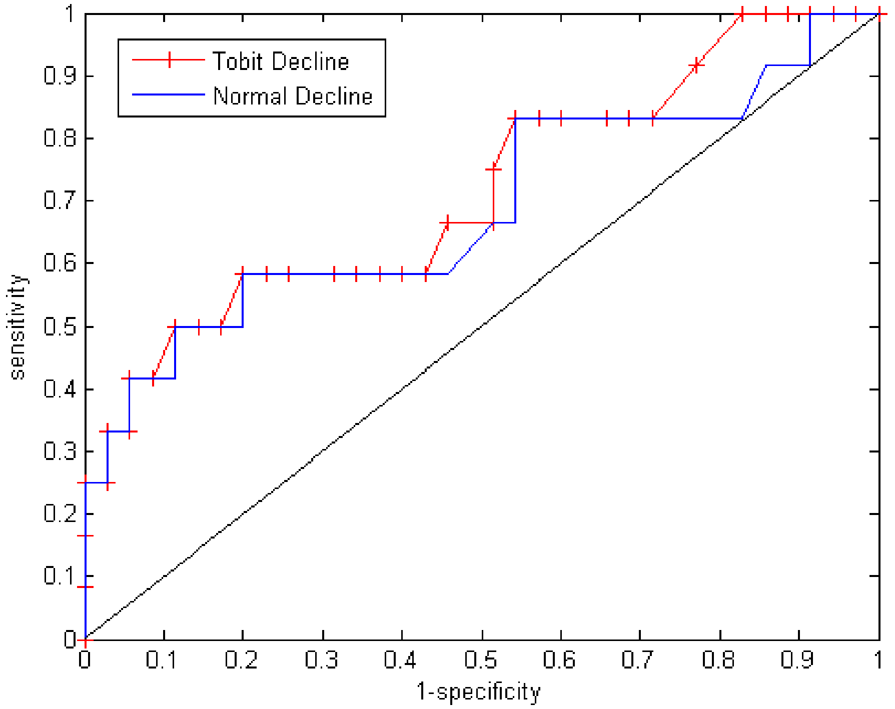

Figure 9 – ROC curve of the quadratic-in-time decline models for the Tobit Decline Model and the normal decline model. The black line is a totally random classifier.

	Linear (Tobit)	Linear (Normal)	Quadratic (Tobit)	Quadratic (Normal)
AUC	.70	.68	.73	.69
$Spec_{80}$	.458	.438	.457	.457
$MSE_{80}(y_5)$	9.81	8.52	7.95	8.05

Table 5 – Classification and score prediction results on the BNT30. AUC is area under the ROC curve, Spec80 is the specificity when sensitivity is .80, and MSE80 is the mean squared error of score prediction when for the classification model with a sensitivity of .80.

The results in Table 5 indicate that the best model with respect to area under the ROC curve was the Tobit Decline Model where the MCI group declined as a quadratic function of time since baseline. This model also resulted in the smallest mean squared error of the estimates of the score at the fifth time point.

6. Application to Test with Floor and Ceiling Effects: Delayed Word Recall

In addition to data with extreme ceiling effects, it is also valuable to compare the performance of censored normal models with standard normal models on empirical data both floor and ceiling effects. The floors also lead to biased estimates of the model parameters, but tests that exhibit both floor and ceiling effects also have the advantage of covering a wider range of measurement than tests with just significant ceiling effects. The test that I will be using that has both floor and ceiling effects is the CERAD 10-Word List Delayed Recall (WLDR). The WLDR is a free recall test where subjects are presented with a list of 10 words that they then have 3 immediate recall trials with, then are distracted by a story task and are then asked to once again recall the list. The score is the sum of unique words from the list that are correctly recalled.

For this work, I am again only considering subjects who were not cognitively impaired during enrollment in the study and who have at least three follow-up visits in which a WLDR score was recorded. Mild cognitive impairment (MCI) is defined as two consecutive visits with a CDR of 0.5 or greater, consistent with how it was defined for the 30-item Boston Naming Test (BNT30) data. For the WLDR data, 565 subjects meet the inclusion criteria, 144 of which have developed MCI. Figure 10 shows the distributions of WLDR scores at baseline and again at the fourth visit. A distinct shift towards the floor is evident.

6.1 Decline Trajectories with Ceiling and Floor

6.1.1 Methods

The mixed model employed for this analysis is the same form (equation 14) as for the BNT30. Models employing linear or quadratic decline with respect to time for the MCI population are again being compared. For the WLDR test, the ceiling C=10 and the floor F=0. The subject-specific coefficients $\mathbf{u}_i$ are assumed to be distributed mean 0 with covariance matrix **Σ**, and ε is assumed to be distributed N~(0,σ^2). Fitting the models in WinBUGS, the prior distributions of the parameters were β_0, β_1, $\beta_{2,}$ β_3~N(0,1E-6) and 1/σ^2~Γ(1,1). The prior distribution of **Σ** is ~Wishart(**I**,4). For each model 1000 samples used for burn-in and the next 1000 samples were taken as the parameter estimates.

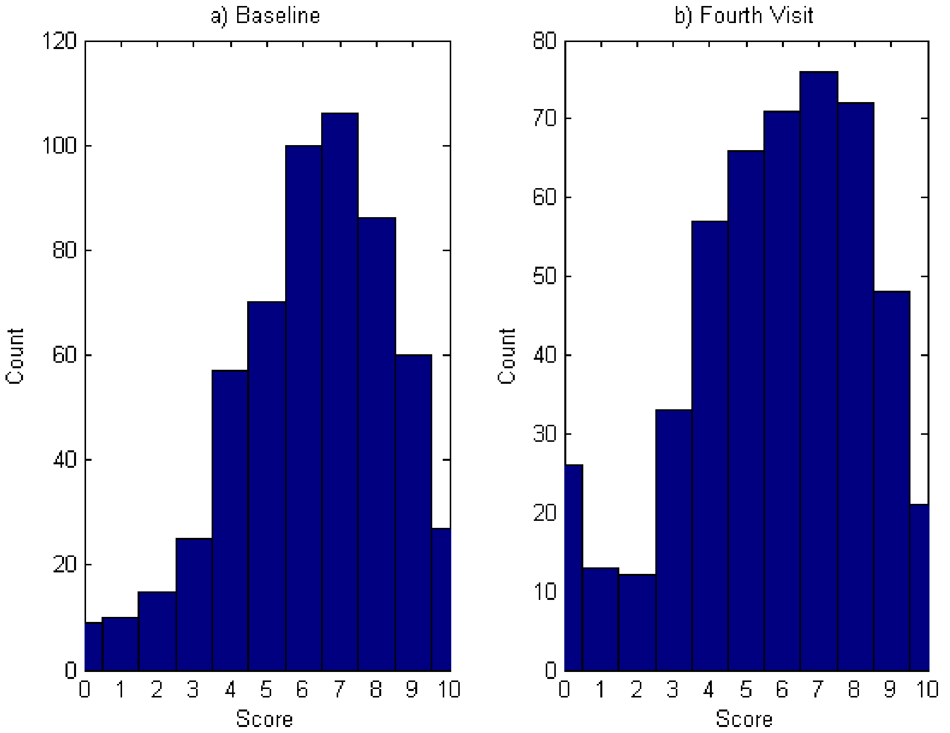

Figure 10 – Histograms demonstrating the distribution of WLDR score at baseline and at the third follow-up, showing a shift of subjects towards the floor of the test.

6.1.2 Results

The results of the fits for the quadratic and the linear decline models are listed in Table 6. If the 95% confidence interval for the parameter estimate includes 0, the parameter is negligible and dropped from the model. Unlike with the BNT30, the WLDR has post-baseline time dependent decline for healthy subjects as well as subjects who develop MCI. This is likely because the WLDR is a more difficult test and the distribution of scores is across the entire range of possibilities.

	Linear (Tobit)				Linear (Normal)				Quadratic (Tobit)				Quadratic (Normal)			
DIC	7436				8167				7434				8174			
β_0	11.0(1.0)				11.2(.6)				10.0(1.0)				11.2(.6)			
β_1	-.057(.012)				-.058(.008)				-.045(.013)				-.058(.008)			
β_2	-.093(.029)				-.098(.029)				-.099(.028)				-.10(.03)			
β_3	-.30(.06)				-.27(.06)				-.095(.024)				-.083(.022)			
σ	1.3(.03)				1.3(.02)				1.3(.03)				1.3(.02)			
Σ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	.09 (.02)	0	0	0	.08 (.02)	0	0	0	.09 (.02)	0	0	0	.08 (.02)	0
	0	0	0	0	0	0	0	.13 (.04)	0	0	0	.04 (.008)	0	0	0	.03 (.006)

Table 6 – Parameters estimates of the linear and quadratic mixed models on the 565 subjects with WLDR data. Values in the () are standard deviations of the estimates.

When accounting for the parameters that are dropped due to insignificance, the model becomes:

$$y_{ij}^{*} = \beta_0 + \beta_0 x_{1i} + (u_{2i} + \beta_2)x_{2ij} + MCI(u_{3i} + \beta_3)x_{3ij} + \varepsilon_i \quad (17)$$

As was the case with the BNT30 scores, the Tobit Decline Model has a lower DIC and thus a higher likelihood than the associated standard normal model, as well as a slightly higher average time dependent decline for the MCI group.

6.2 Score Prediction and Identification of Mild Cognitive Impairment

6.2.1 Methods

In the same way as was done with BNT30 data, classification and prediction are evaluated together for the WLDR data. Classification is done using the log-likelihood ratio (equation 16) of the group classification conditioned on the regressed score. Regressions are done against the scores of a training set to determine the probability distributions for MCI and healthy subjects given score estimates exactly as was done with the BNT30 data in section 5.2. The priors are taken as the proportions of MCI and healthy subjects from the training set, and the classification threshold is varied so a Receiver Operator Characteristic (ROC) curve can be constructed. For this analysis, the training set again consisted of the subjects who only completed four score evaluations (N_{tr}=332, N_{trMCI}=49). The evaluation set consisted of subjects who had completed at least five score evaluations, although it was only evaluated on their fifth (N_{ev}=233, N_{evMCI}=95).

To evaluate score prediction, the evaluation set's subject scores are predicted for their fifth evaluation which occurs at time t_{i5} after baseline and the predicted score is compared to their actual score. The predicted score value depends on the parameters of the model, which is in turn dependent on the classification. The score predictions are evaluated by determining the mean squared error of the score fifth evaluation score predictions under the prior probability distribution resulting in 80% sensitivity to MCI.

6.2.2 Results

ROC curves for the classification under the linear and quadratic models from section 6.1 are shown in Figures 11 and 12, respectively. The areas under the ROC curve (reported as AUC) as well as the specificity of the model when the priors are chosen for 80% sensitivity to prediction of MCI as a positive are reported in Table 7.

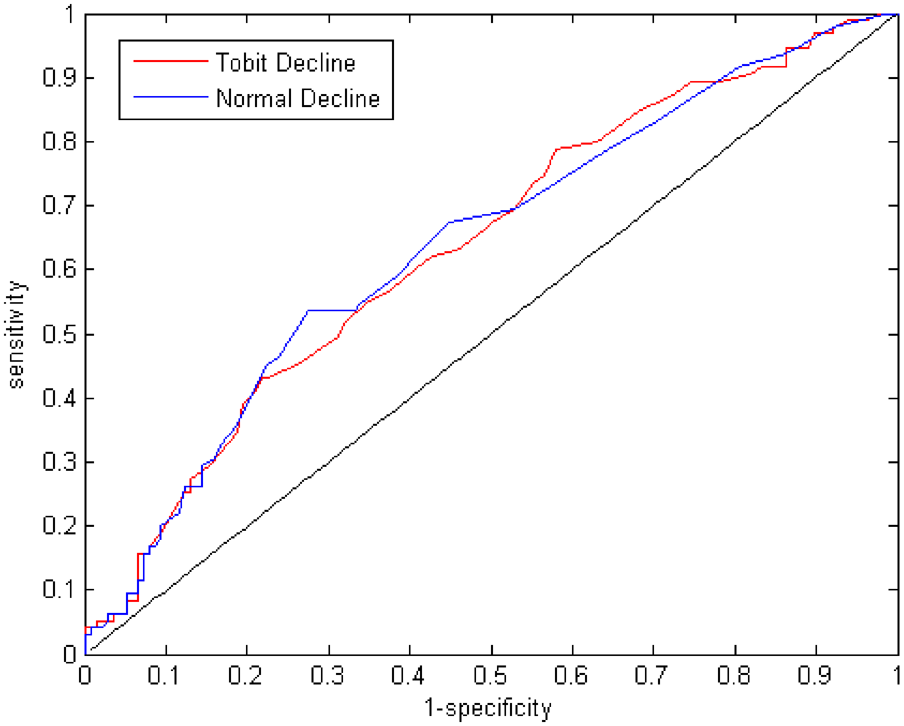

Figure 11 – ROC curve of the linear-in-time decline models for the Tobit Decline Model and the normal decline model. The black line is a totally random classifier.

	Linear (Tobit)	Linear (Normal)	Quadratic (Tobit)	Quadratic (Normal)
AUC	.645	.657	.684	.629
Spec_{80}	.370	.078	.457	.203
$\text{MSE}_{80}(y_5)$	4.87	5.33	4.95	5.35

Table 7 – Classification and score prediction results on the BNT30. AUC is area under the ROC curve, Spec80 is the specificity when sensitivity is .8, and MSE80 is the mean squared error of score prediction when for the classification model with a sensitivity of .8.

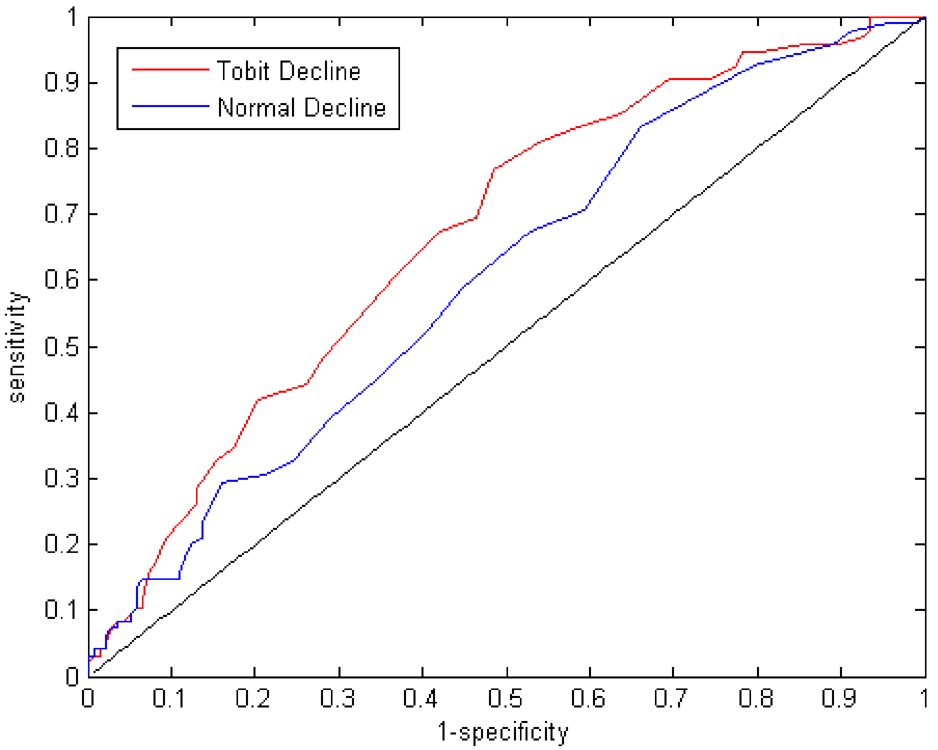

Figure 12 – ROC curve of the quadratic-in-time decline models for the Tobit Decline Model and the normal decline model. The black line is a totally random classifier.

The results in Table 7 indicate that the best model with respect to area under the ROC curve was the Tobit Decline Model where the MCI group declined as a quadratic function of time since baseline (also quite apparent in Figure 12). However, the Tobit Decline Model with a linear time-dependent decline resulted in a slightly smaller mean squared error for the prediction of the fifth score evaluation. Both the linear and quadratic Tobit Decline Models resulted in lower mean squared errors than the normal decline models.

7. Dropout under Relative Performance Constraint

7.1 Methods

While early estimation of cognitive decline is confounded by ceiling effects in data, scores several years after baseline might be overestimated if low performers are dropping out and the remaining higher scorers are assumed to be distributed in the same manner as the entire population. To explore the possibility of overestimating scores when not accounting for informative dropout, we simulated a population of 200 subjects that had scores initially distributed y~N(40,16) at age 60. Scores were declined linearly at an average rate of 0.5, with errors at each time point distributed ε_t~N(0,4). At age 72, the bottom 7.5% of the remaining population was dropped at each followup.

To fit the data, the assumptions were that subject relative performance within the population remained constant, and that the distribution of scores was Gaussian (these are the conditions under which the data was generated). At baseline, we determine the relative performance of each subject by estimating μ and σ of the baseline distribution and then calculating

$$p_i = \frac{1}{2}\left[erf\left(\frac{y_{it} - \mu_t}{\sigma_t\sqrt{2}}\right)\right] \tag{18}$$

Then, at each time point after baseline, estimate μ and σ from the subjects who have not dropped out, using the relative performance at the previous time point to generate an estimate of the population parameters from each subject. Then the estimate of μ and σ for that time point is the average of the estimates across remaining subjects. Estimates of missing subject data can then be generated using

$$\hat{y}_{it} = \hat{\mu}_t + \hat{\sigma}_t \sqrt{2} erf^{-1}(2p_{it-} - 1) \qquad (19)$$

7.2 Results

Results of the population estimates compared to true values for a variety of fits are given in Figure 13.

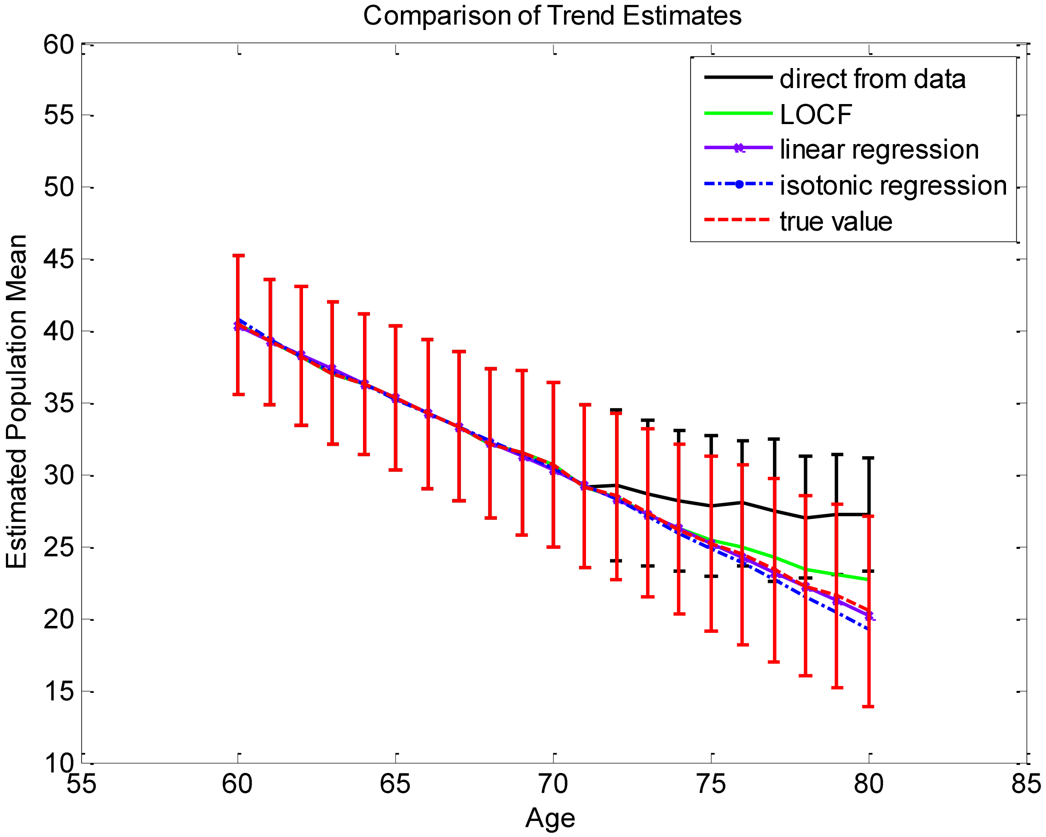

Figure 13 – True population parameters versus estimates obtained directly from data, using last observation carried forward, a linear regression across means, and a monotonic regression.

As is apparent in Figure 13; estimating population parameters directly from the remaining subject data results in an overestimation of the mean score, as well as an underestimation of the variance. Using last observation carried forward (LOCF) improves population estimates, but provides poor estimates of missing individual data and still overestimates the mean. Linear and isotonic fits to the mean using the above estimate methods of relative performance constraint for recovering missing data recovered population estimates well (max error 1.22 on population mean, mean square error of 0.32). An important note on these results is that the relative performance constraint may not hold true. Subjects who begin to develop dementia should be expected to drop relative to the population that they were classified with when they were healthy, in which case this method could be adopted to check for outliers as a possible method of detecting cognitive impairment in future work.

8. Discussion and Future Directions

Simulation results demonstrate that when observable data is normally distributed with a mean dependent linearly on covariates and a fixed standard deviation, but censored such that all values above a ceiling C are observed as being C, fitting a normal linear regression model without accounting for the ceiling results in miss-estimation of the model parameters (Section 4.1). This results in an underestimation of the mean and variance of the observable score data (y). Fitting the data with a censored normal model recovers the true distribution parameters with high confidence.

Furthermore, simulation results demonstrate that when observable outcomes in a declining population are normally distributed with a significant portion censored at ceiling C such that they are observed as C, the rate of decline estimated by a normal linear decline model is lower than the actual rate of decline. This results in a lower estimate of the mean of the observables than a model that accounts for the censoring (Section 4.2). As time progresses from the initial observation point the estimate of the observable population by a normal decline model that does not incorporate censoring improves. This is because the effect of the censoring becomes less significant as the data declines, as a progressively smaller proportion of the observables are cut off by the ceiling.

Simulation results indicate that in noisy data under piecewise linear decline with a changepoint τ, when the observable data is censored such that values above a ceiling C are observed as C, the estimated location of that changepoint under a normal changepoint model is poorly identified. Applying a censored normal model resulted in a more accurate estimation of the changepoint (Section 4.3), although both estimation methods performed poorly.

When decline models are applied to empirical data with significant ceiling effects at baseline such as the 30-item Boston Naming Test (BNT30) data, the likelihood of the data under the Tobit Decline Models exceeds that of a normal decline model (Section 5.1). Furthermore, when evaluating the classification accuracy of normal versus Tobit Decline Models, the area under the ROC curve for the Tobit Decline Models was lower than the normal decline models (Section 5.2). The mean squared error of future score estimates is lowest when adopting a model where the decline in BNT30 score is dependent on the square of time since baseline for mild cognitive impairment (MCI) subjects. For decline models applied to empirical data with both ceiling and floor effects at baseline such as the Word List Delayed Recall (WLDR) data, the likelihood of the data under the Tobit Decline Models exceeds that of a normal decline model (Section 6.1). When evaluating the classification accuracy of normal versus Tobit Decline Models, the area under the ROC curve for the Tobit Decline Model in WLDR scores was lower than the normal linear mixed model when the time-dependant change for MCI subjects is quadratic (Section 6.2). Tobit Decline Models with MCI subjects undergoing a quadratic time-dependent decline had a lower mean squared error in predicted score values than normal linear mixed models.

While early estimation of decline can be underestimated in high-performing subjects due to the within-test ceiling and floor effects, dropout patterns that are a result of poor performance can lead to an overestimation of population performance if not accounted for. Under a model where subject relative performance within a population remains consistent, simulation results (Section 7) indicate that modeling longitudinal decline as a constant relative performance under changing population parameters recovers individual scores with less error than standard fitting methods, as well as that not accounting for dropout overestimates performance at higher ages. In future work, it will be important to account for dropout if longer longitudinal studies are used. If the assumption of consistent relative performance to the rest of the population over time does not hold, additional constraints will need to be included in the models. It may also be possible to utilize the change in relative performance to suggest a transition from healthy to MCI.

While the results of the work presented here indicate that Tobit Decline Models applied to longitudinal data with ceiling and floor effects provide slightly improved classification of MCI and prediction of future score estimates, applying such models to neuropsychological tests with a fixed number of items such as the BNT30 or the WLDR is not entirely accurate. While it is possible that the distributions of scores for longer versions of each test could be normal, the very act of extending the length of the test would change the distribution parameters such that they would not be the same as they are when fixed at length C (the same C that is the ceiling in the Tobit Decline Models). A sounder model would be to model test performance as a sum of Bernoulli distributions where the probability of a correct score on item k of a test is modeled as

$$p_{\mathrm{k}} = f(s;\theta_p) \quad \begin{cases} p_k \to 1\ as\ s_i \to \infty \\ p_k \to 0\ as\ s_i \to \ \ 0 \end{cases} \tag{20}$$

where s is a measure of subject cognitive ability for measured by the test, which follows an assumed distribution with good face validity (such as lognormal for memory capacity in free recall ability for the WLDR test). The parameters θ_p are fit under the constraints of the model to reflect the difficulty of the items, the minimum chance of correct answer, and the rate at which chance changes based on subject ability s.

The overall score on a test can then be modeled as an average over the item-level model, where the estimates of the item-level parameters can be calculated from experiments where performance on each item presented are recorded. The probability across items can be average and a model for total score can be estimated as a binomial distribution with n items and probability p given by the average over the item-level model with the understanding that the overall score variance will be underestimated item-level models. Furthermore, total-score models cannot identify items that are more discriminatory than others.

References

[1] R. Savica and R. C. Petersen, "Prevention of Dementia," *The Psychiatric Clinics of North America,* vol. 34, pp. 127-145, 2011.

[2] F. Portet, P. J. Ousset, P. J. Visser, G. B. Frisoni, F. Nobili, P. Scheltens, B. Vellas and J. Touchon, "Mild cognitive impairment (MCI) in medical practice: a critical review of the concept and new diagnostic procedure. Report of the MCI Working Group of the European Consortium on Alzheimer's Disease," *Journal of Neurology, Neurosurgery & Psychiatry,* vol. 77, pp. 714-718, 2006.

[3] R. C. Petersen, "Mild cognitive impairment as a diagnostic entity," *Journal of Internal Medicine,* vol. 256, pp. 183-194, 2004.

[4] J. Verghese, M. Robbins, R. Holtzer, M. Zimmerman, C. Wang, X. Xue and R. B. Lipton, "Gait Dysfunction in Mild Cognitive Impairment Syndromes," *Journal of the American Geriatrics Society,* vol. 56, pp. 1244-1251, 2008.

[5] R. Camicioli, D. Howieson, B. Oken, G. Sexton and J. Kaye, "Motor slowing precedes cognitive impairment in the oldest old," *Neurology,* vol. 50, pp. 1496-1498, 1998.

[6] P. Rabbitt, M. Lunn, D. Wong and M. Cobain, "Age and Ability Affect Practice Gains in Longitudinal Studies of Cognitive Change," *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences,* vol. 63, pp. P235-P240, 2008.

[7] H. Jacqmin-Gadda, D. Commenges and J. Dartigues, "Random Changepoint Model for Joint Modeling of Cognitive Decline and Dementia," *Biometrics,* vol. 62, pp. 254-260, 2006.

[8] D. B. Howieson, N. E. Carlson, M. M. Moore, D. Wasserman, C. D. Abendroth, J. Payne-Murphy and J. A. Kaye, "Trajectory of mild cognitive impairment onset," *Journal of the International Neuropsychological Society,* vol. 14, pp. 192-198, 2008.

[9] A. E. Zehnder, S. Blasi, M. Berres, R. Spiegel and A. U. Monsch, "Lack of Practice Effects on Neuropsychological Tests as Early Cognitive Markers of Alzheimer Disease?" *American Journal of Alzheimer's Disease and Other Dementias,* vol. 22, pp. 416-426, 2007.

[10] D. L. Streiner, "The Case of the Missing Data: Methods of Dealing With Dropouts and Other Research Vagaries," *Can J Psychiatry,* vol. 47, pp. 68-75, 2002.

[11] F. J. M. M. Molnar, B. M. Hutton and D. M. H. A. P. Fergusson, "Does analysis using "last observation carried forward" introduce bias in dementia research?" *Canadian Medical Association Journal,* vol. 179, pp. 751-753, 2008.

[12] G. Yi and W. He, "Median Regression Models for Longitudinal Data with Dropouts," *Biometrics,* vol. 65, pp. 618-625, 2009.

[13] L. Su and J. W. Hogan, "Varying-coefficient models for longitudinal processes with continuous-time informative dropout," *Biostatistics,* pp. kxp040, 2009.

[14] Y. Ying and J. A. L. Roderick, "Mixed-Effect Hybrid Models for Longitudinal Data with Nonignorable Dropout," *Biometrics,* vol. 65, pp. 478-486, 2009.

[15] D. Hedeker and R. D. Gibbons, "Application of Random-Effects Pattern-Mixture Models for Missing Data in Longitudinal Studies," *Psychological Methods,* vol. 2, pp. 64-78, 1997.

[16] H. Dodge, C. Shen and M. Ganguli, "Application of the Pattern-Mixture Latent Trajectory Model in an Epidemiological Study with Non-Ignorable Missingness," *J Data Sci,* vol. 6, pp. 247-259, 04, 2008.

[17] L. Wang, Z. Zhang, J. J. McArdle and T. A. Salthouse, "Investigating Ceiling Effects in Longitudinal Data Analysis," *Multivariate Behavioral Research,* vol. 43, pp. 476-496, 2011/08/08, 2008.

[18] B. Uttl, "Measurement of Individual Differences," *Psychological Science,* vol. 16, pp. 460-467, 06/01, 2005.

[19] T. Amemiya, "Regression Analysis when the Dependent Variable Is Truncated Normal," *Econometrica,* vol. 41, pp. 997-1016, 11/01, 1973.

[20] W. Gilks, S. Richardson and D. Spiegelhalter, *{M}Arkov Chain {M}Onte {C}Arlo Methods in Practice.* Chapman and Hall, 1996.

[21] W. J. Mack, D. M. Freed, B. W. Williams and V. W. Henderson, "Boston Naming Test: Shortened Versions for Use in Alzheimer's Disease," *J. Gerontol.,* vol. 47, pp. P154-P158, 05/01, 1992.