The Waiting Game: A Survival Analysis of Unemployment Duration in South Africa 2001-2004
By Kristoff Potgieter PTGKRI001

Abstract:

Who faces the worst labour market prospects in South Africa? To answer this pertinent question I will use panel data from the biannual Labour Force Survey composed by Statistics South Africa, to estimate time spent unemployed by respondents. Survival analysis is then used to untangle the characteristics and determinants of unemployment duration based on several variables characterising the demographic, geographic and educational diversity of the South African labour force. Results from the analysis support a ranking model of unemployment, as proposed by Blanchard and Diamond (1990), with divergent unemployment exit rates between high and low ranked work seekers. The ranking given to a given educational level is found to be strongly related to race, indicating that individuals from non-"model-C" schools face inferior labour market conditions. There is also some evidence that a willingness to work in the informal sector raises the probability of transitioning out of unemployment amongst individuals with less than a completed secondary education.

This thesis was kindly supervised by Professor Murray Leibbrandt at the School of Economics, University of Cape Town. Additional, thanks go to Economics Research South Africa and the National Research Fund for their generous funding of my graduate studies at the University of Cape Town.

Kristoff Potgieter

Kristoff.potgieter@gmail.com

Table of Contents

1 Introduction

The persistence of high levels of unemployment in South Africa remains a source of significant political and social concern, while simultaneously a topic of interest to empirical economists. In this thesis I will focus on one aspect of the South African unemployed, which, up till now, has received limited attention; namely the demographic characteristics of long-term unemployment duration. My analysis will make use of survival analysis methodology. While the analysis will make no causal assumptions in the determination of unemployment duration, it will provide a detailed descriptive analysis of the labour market conditions experienced by various demographic and educational groups within South African society. The overarching theme of this thesis will be to answer the question as to which characteristics are associated with the worst, and best, labour market experiences.

Before starting with the empirical analysis a brief history of the factors driving South African unemployment will be outlined an overview of the theoretical framework of the job-search model. Next the focus turns to the primary aim of this study: an empirical Survival analysis of South African unemployment. Section 2 gives a description of the sourced and derived data used in this study, obtained from the bi-annually collected Labour Force Survey compiled by Statistics South Africa from September 2001 to March 2004. Section 3 contains the empirical survival analysis of the data. Finally, section 4 concludes by summarizing the findings of my empirical analysis and outlining potential areas of research interest.

1.1 The South African Labour Force 1994-2004

A high level of unemployment has been characteristic of South Africa for more than forty years, with unemployment increasing over the 70s, 80s and 90s. Importantly, it has long been true that the unemployed in South Africa tend to remain so for extended periods of time. By the mid 90s more than two thirds of the unemployed in South Africa had never worked for pay (Standing *et al. 1996*). The problem of long-term unemployment has not disappeared in the post-apartheid era; the 2005 LFS indicated that that 59% of the unemployed had never held a job while 40% of the unemployed had been unemployed for a period greater than 3 years (Lam *et al.* 2009).

In the post-apartheid period the South African economy saw a normalization of trade relations, along with a liberalization of the economy. Over the period several initiatives by the South African government have attempted to redress historical inequities, through the expansion of social welfare programs, and improving the international competitiveness of South African industries. Both the goals and strategies of the government over this period were articulated through the medium-term macro- economic plans namely the Growth, Employment and Redistribution Plan

(GEAR), and its successor, the Accelerated and Shared Growth Initiative for South Africa (ASGISA). Despite the optimism of these projects, economic growth did not have a significant effect on job creation (Bhorat & Oosthuizen, 2005). Multiple perspectives exist as to the contemporary sources of unemployment in South Africa (for a survey of the literature see Fourie 2011). While my analysis will be agnostic of these models I will give a brief overview of the major demographic and legal changes observed in the post-apartheid labour market before continuing.

In their paper on the determinants of South African unemployment Banarjee *et al.* (2008) suggest that several structural changes have increased the equilibrium unemployment rate in South Africa in the post-apartheid era. These factors consisted of a mixture of demographic, legal and technological changes, the shifting of the political-economic paradigm and the remnants of apartheid-era policies.

Table 1.1 Trends in the compositional change of the South African Labour Force 1995- 2002

	% of growth in the labour force	% of labour force 2002
Females	61.4%	49.3%
Africans	85.2%	75.5%
Urban	59.8%	62.5%
No matric	55.2%	63.7%
Under 35	55.0%	54.1%

Source: Bhorat & Oosthuizen 2005 from OHS 1995 and the LFS 2002

Firstly, there has been a significant growth in labour market participation. The period from 1995 to 2002 saw an increase of 5.2 million individuals in the South African labour force. This growth in the labour force outstripped job creation, limiting the ability of South African economy to meet the targets set in terms of reducing the unemployment rate. Growth in the South African labour force was disproportionately distributed amongst the various demographic groupings with the majority of growth attributable to Africans, females, urban dwellers, youths and individuals with incomplete secondary educations (Table 1.1). The links between demographic and educational characteristics and unemployment duration will form the bulk of my analysis.

Secondly, skill biased technical changes in the economy led to diminishing labour demand in traditionally low skill intensive sectors such as mining and agriculture. These changes have are likely to have exacerbated labour market differences between skilled and unskilled workers (Banarjee *et al.* 2008).

Thirdly, there has remained a low level of political will in increasing international labour market competitiveness by devaluing the currency or introducing more flexible labour regulations. This has been accompanied by increasing political strength amongst labour-unions which have steadily increased the wage premium of union members. The combination of restrictive labour regulations and high levels of union power has given rise to a large informal sector, where employers can avoid compliance with labour laws and pay lower wages (Fourie, 2011).

Finally, African work seekers have faced considerable search costs due to spatial labour market distortions and other vestiges of racial discrimination caused by the apartheid era policies. In later sections of this paper I will investigate the effects of geographical location and race on time spent unemployed.

The combination of these factors have created an unemployment rate of 37%. Many of the unemployed have never held formal work and live in abject poverty, which greatly limit their search efforts (Kingdon and Knight, 2004). While these discouraged work seekers are not included in the official definition of unemployment they constitute 12% of the total labour force and nearly a third of the unemployed (Fourie, 2011). Discouraged work seekers were included amongst the unemployed for my analysis.

1.2 The Search Model

The theoretical framework used in this study is that of the job search model as treated by Mortensen (1986). The use of the search model analysis of the labour market stems from the empirical findings, which suggested that differences in unemployment rates between demographic groups can largely be accounted for by differing frequency and duration of employment spells. Consequently, economists started viewing a workers labour market history as the realization of a stochastic process involving individual decisions, which could potentially be modeled as a Markov-process (Mortenson, 1986). This section will offer a brief overview of the model as it applies to the empirical analysis in section 4. For further insight into the origins and theoretical framework of the search model see Stigler (1961, 1962), Gronau (1971), McCall (1970) and Mortenson (1970).

Within the search framework individuals enter the labour market and effectively "shop" for a job. An individual will expend both pecuniary resources along with time and effort to obtain the optimal job potentially available to them. The job offers received by an individual are not perfectly related to the effort expended thereupon but may vary according to several external, and potentially unknown, factors such as job availability or employer preferences. Within this framework of uncertainty an individual will choose the optimal level of expenditure to be used in the job search in order to maximize discounted future utility. To simplify the analysis I will use the most basic form of the standard search model, the stationary search model.

Let us assume that, over a given period t, an individual faces discount rate beta(t), and that she receives n job offers with probability N, where N has a probability density function q(n, h). Additionally, wages follow a cumulative density function F(w). Job seekers enter the labour market with a given reservation wage, any job offers with wages below the reservation wage limit will not be accepted. Theoretically the reservation wage is expected to decrease, as time goes by, in order to increase to probability of encountering an acceptable job offer. Once a job seeker receives an acceptable offer she transitions out of unemployment into employment.

Importantly for my analysis the demographic factors can potentially affect either the rate of job offers or the job acceptance rate. People with higher educational levels, for example, may be expected to receive a greater number of job offers while simultaneously being less likely to accept low wage offers. For my purposes I will make no distinction between factors which affect the job offer rate or those which affect the reservation wage, and thereby the job acceptance rate.

A further limitation of my analysis is that it treats all unemployed individuals equally once the relevant demographic and educational factors have been accounted for. That is to say I assume the search function to be a stationary process, independent of the labour market participation before the first recorded date of unemployment.

The framework outlined above is purposefully simplistic and brief. However, as I wish to focus solely on a single transition into employment for each individual it is suitable for my purposes. Through the matching process it is assumed that an economy will approach its natural or equilibrium rate of unemployment. The process whereby the unemployed find work is, however, likely to vary greatly. In later sections of this paper I will consider the possibility of ranking in the South African labour force as hypothesized in the model of Blanchard and Diamond (1990).

2 Overview of the Data

The data used in this paper was sourced from the panel data of Labour Force Survey (LFS) and comprises six waves of a biannual survey, collected by Statistics South Africa (StatsSA) over the period from September 2001 to March 2004. The panel identifiers of the data set will allow us to follow individuals over time and extrapolate labour market transitions. This is not the first analysis to make use of the LFS, in their study of labour market transitions Dinkelman and Ranchod (2008) give a full overview of the construction and composition of the dataset. Originally the LFS was intended for use as a barometer of the South African labour force, assuming the role previously played by the October Household Survey. As such the data collected has the benefit of being both representative of the South African labour force while simultaneously being quite rich in terms of the variables included therein. Despite these benefits the LFS was not originally intended for use in duration analysis and several difficulties arise when using it in this manner.

The panel data used for the construction of the unemployment duration data was not a static panel, which follows the same individuals over all the waves of the panel. Instead, a rolling panel was employed replacing 20% of the respondents from the previous waves with new ones. Respondents who changed residence between panels were also lost from the panel (Dinkelman and Ranchod, 2008). This is potentially problematic as movement related to the outcome variable of interest, such as would occur if individuals move to find work, could lead to biased results in my analysis. For the purpose of this study I will assume that attrition from the panel was at random. Additionally, the large gaps between waves of the panel limits the scope of my analysis to long-term unemployment.

2.1 Demographic Data

Table 2.1 Demographic Characteristics of the Panel

Date	Sept 2001	Feb 2002	Sept 2002	Mar 2003	Sept 2001	Mar 2004	Total
Male	28,377	33,349	31,897	31,409	27,461	21,653	**174,146**
Female	32,262	37,806	36,154	35,417	31,294	24,201	**197,134**
Unspecified	0	0	0	0	2	5	**7**
Less than 15	20,269	23,876	23,090	21,526	18,212	12,568	**119,541**
15-30	17,675	20,630	19,642	19,521	17,286	13,773	**108,527**
31-65	19,755	23,179	21,871	22,268	20,037	16,849	**123,959**
More than 65	2,940	3,470	3,448	3,511	3,222	2,669	**19,260**
Unspecified	297	363	343	336	317	271	**1,927**
Western Cape	6,356	7,664	7,445	7,573	6,300	5,208	**40,546**
Eastern Cape	9,173	10,843	9,951	10,039	8,543	6,776	**55,325**
Northern Cape	2,898	3,630	3,362	3,171	2,887	2,227	**18,175**
Free State	4,975	5,708	5,309	5,173	4,973	3,699	**29,837**
KwaZulu Natal	10,820	11,924	10,646	10,898	9,166	7,498	**60,952**
North West	6,599	8,022	7,369	6,950	6,011	4,455	**39,406**
Gauteng	6,528	8,368	7,871	7,505	6,816	5,232	**42,320**
Mpumalanga	5,403	6,425	6,783	6,498	6,076	4,453	**35,638**
Limpopo	7,887	8,571	9,315	9,019	7,985	6,311	**49,088**
African	47,659	54,802	52,681	51,304	45,143	34,939	**286,528**
Coloured	7,287	8,721	8,179	8,234	7,073	5,570	**45,064**
Indian/Asian	1,508	1,733	1,540	1,624	1,304	1,063	**8,772**
White	4,185	5,899	5,651	5,661	5,232	4,282	**30,910**
Unspecified	0	0	0	3	5	5	**13**
Rural	28,106	31,802	31,075	30,258	26,803	21,162	**169,206**
Urban	32,533	39,353	36,976	36,568	31,954	24,697	**202,081**
Total	**60,639**	**71,155**	**68,051**	**66,826**	**58,757**	**45,859**	**371,287**

The composition of the rolling panel makes it difficult to give an accurate overview of the entire panel. For this reason descriptive statistics are given on both a sample total and a wave-by-wave basis. Demographically the data collected covers all strata of South African society as represented by gender, age, race and geographical location. Additionally the coding of the unique person identifiers allowed for the extraction of a rural-urban indicator variable. As is to be expected in a representative sample, the majority of the sample consists of working age Africans. Non-response does not seem to be a problem, as the vast majority of respondents have all demographic data recorded.

Table 2.2 Educational Attainment in the Panel

Date	Sept 2001	Feb 2002	Sept 2002	Mar 2003	Sept 2001	Mar 2004	Total
No-High school	35,483	40,150	37,918	35,799	30,843	22,686	**202,879**
Some High School	15,561	18,466	17,743	17,986	15,743	12,696	**98,195**
Matric	6,101	8,039	6,941	6,912	6,302	5,475	**39,770**
NTC	636	778	1,930	2,406	2,516	2,232	**10,498**
Diploma/Certificate	1,806	2,301	1,843	1,838	1,551	1,219	**10,558**
University Degree	991	1,370	1,544	1,732	1,592	1,345	**8,574**
Other/Don´t know	61	51	132	153	210	206	**813**
	39,568	**53,847**	**48,829**	**45,668**	**43,890**	**29,344**	**261,146**

Table 2.2 above concludes my overview of the panel by summarizing the educational levels of those surveyed. Once again I note significant variation across the panel. As expected, those with lower levels of education predominate in the sample. Survey non-response was not a problem with regards to the education variable.

Table 2.3 Provincial Unemployment rates over the sample period

Date	Sept 2001	Feb 2002	Sept 2002	Mar 2003	Sept 2001	Mar 2004
Western Cape	24.70%	24.18%	25.69%	25.14%	25.22%	23.38%
Eastern Cape	42.88%	34.12%	43.32%	44.58%	45.39%	48.26%
Northern Cape	31.85%	39.41%	28.98%	33.19%	31.61%	34.13%
Free State	30.13%	31.74%	31.66%	33.95%	37.88%	31.82%
KwaZulu Natal	36.45%	38.79%	34.82%	37.38%	34.31%	35.66%
North West	40.98%	36.21%	40.72%	38.52%	40.96%	35.90%
Gauteng	30.68%	34.62%	30.96%	35.13%	30.67%	32.84%
Mpumalanga	31.02%	31.05%	30.93%	34.09%	31.11%	32.43%
Limpopo	47.29%	49.11%	49.51%	54.83%	49.02%	50.75%

Source: StatsSA 2008;2009

To estimate accurately of the characteristics of unemployment duration I will include an measure of the local unemployment rate faced by individuals in the sample. Given South Africa´s geographical size, along with large-scale heterogeneity in economic activity between provinces, it seems likely that unemployment rates should differ substantially, both between locations and within identical locations over time.

Some question may arise as to which unemployment rate to use for analysis. Ideally one would wish to collect local unemployment rates by sampling cluster for each period. This would present a highly

accurate snapshot of employment conditions facing a given individual. However, no external source exists to estimate such unemployment rates and sampling clusters were not identified in the data collected by StatsSA. Individuals were therefore assigned with the regional unemployment rate of their province for each period, using the expanded definition of unemployment. While some accuracy is lost in this way the combination of provincial unemployment rates along with an individual´s urban-rural status should capture most geographical variation in labour market matching. Data on provincial unemployment rates was compiled from revised versions of the LFS (StatsSA 2008; 2009).

2.2 Labour Market Transitions

Table 2.4 Employment Status and Sectoral Composition in the Panel

	Non-Continuous	Continuous	Total
2 Observations	13,059	36,062	**49,121**
3 Observations	15,843	23,876	**39,719**
4 Observations	7,001	10,584	**17,585**
5 Observations	2,482	10,200	**12,682**
6 Observations	0	3,356	**3,356**
Total	**38,385**	**84,078**	**122,463**

Essential for the survival analysis performed in the following section of this paper, the data is represents individuals of all possible labour market statuses. Those characterized as employed include workers in the formal and informal sectors of the economy.

Table 2.5 Characteristics of Matched Observations in the Panel

	Sept 2001	Feb 2002	Sept 2002	Mar 2003	Sept 2001	Mar 2004	Total
Employed to Unemployed	.	1,272	1,380	997	977	958	5,584
Unemployed to Employed	.	1,403	1,326	1,090	1,061	1,028	5,908
Total		**2,675**	**2,706**	**2,087**	**2,038**	**1,986**	**11,492**
Employed to economically inactive	.	801	1,047	745	661	753	4,007
Economically inactive to employed	.	1,072	703	643	538	678	3,634
Total	.	**1,873**	**1,750**	**1,388**	**1,199**	**1,431**	**7,641**
Unemployed to Economically inactive	.	1,049	1,154	1,017	1,011	1,113	5,344
Economically inactive to unemployed	.	1,810	1,266	1,421	1,067	1,223	6,787
Total	.	**2,859**	**2,420**	**2,438**	**2,078**	**2,336**	**12,131**
Entered unemployed	9,595	3,801	1,503	2,464	2,038	.	19,401

In the data set compiled StatsSA included all individuals matched for at least 2 observations. An individual could be matched for two or more waves of the panel and could be followed either continuously, without missing observations between her first and last observation, or non-continuously, where some observations are missing between the first and last recorded observation. Note that a total of 122 463 individuals where followed over the span of the of the panel the majority of whom were followed continuously for a period of 2 or 3 waves, or 365 to 548 days of analysis time.

Table 2.6 Transitions between Employment Classifications

Group by type of shift	Period First at Risk	Period Failed/Censored
Employed to Unemployed	Last recorded date of employment.	Fails in period return to employment or right censored at last recorded observation.
Unemployed to economically inactive	Last recorded date of employment.	Censored when shifting to economically inactive.
Economically inactive to unemployed	At risk from first period entering unemployment.	Fails in period return to employment or right censored at last recorded observation.
Economically inactive to employed	Last period recorded as economically inactive/last re	Instantaneously fails during first period of unemployment.
Entered unemployed	Last recorded date of employment.	Fails in period return to employment or right censored at last recorded observation.

Table 2.6 examines recorded shifts in economic status observed in the panel along with the amount of individuals who entered the sample as unemployed. The sample shows significant mobility between states, indicating the possibility of completing duration analysis for a given state.

Table 2.7 Measuring Unemployment Duration

	Before Censoring		After Censoring	
Category	**total**	**mean**	**total**	**mean**
no. of subjects	12334		12334	
no. of records	23148	1.876763	21881	1.774039
(first) entry time		0		0
(final) exit time		544.171		500.1035
subjects with gap	0		0	
time at risk	6711805	544.171	6168276	500.1035
failures	5039	0.4085455	4206	0.3410086

For the use of unemployment duration analysis I need to assemble a data set for all periods individuals wherein individuals could have been considered unemployed or job seekers. Above I have summarized which observations in the original dataset are potentially usable for duration analysis. Note that the period last worked was estimated using a mixture of shifts in individual status over waves of the panel along with individual responses to a question on the survey regarding the last period worked. Where individuals entered the panel as unemployed the period last employed was calculated as the date of entry into the panel subtracted by the period the responded stated as being unemployed. Periods of unemployment were taken as midpoints of the potential categorical responses in the survey, and given an upper limit of 3.5 years for individuals who stated they were unemployed for more than 3 years. Where individuals entered the data set unemployed without indicating their last period employed they were eliminated from the data set to limit bias on estimated unemployment duration.

If respondents shifted from being employed to unemployed between two waves of the panel period last employed was coded as above, if the survey indicated that the last period worked was after the date of the last survey. For all other individuals transitioning from employment to unemployment the date last employed was recorded as that of the last wave of the survey.

Non-economically active individuals are excluded from the analysis until they are either classified as employed or unemployed. Where transitions occur directly from economic inactivity to employment, unemployment duration is calculated from the last period not-economically-active. This contradicts the methodology of Flinn and Heckman (1982) who coded non- participating individuals as unemployed. Notably however, the study completed by Flinn and Heckman focused on sample of 20-24 year old white high school graduates whereas in the LFS individuals could have been coded as non-economically active for reasons as diverse as being enrolled in high school or being pregnant.

The diverse nature of the data set presented here differs starkly from several studies of unemployment duration, which have tended to focus on a relatively small subset of the population (see for example Flinn and Heckman 1982; or Blau 1992). These studies limited the focus of their data sets, in order to ease estimation of the effects of specific policies, by eliminating potentially conflagrating factors. In my study I will not limit my analysis in a similar fashion for two reasons. Firstly, the demographic characteristics of the sample, such as racial segmentation, urban-rural differences and educational attainment in the labour market, are of special interest in the South African case. Secondly, the long duration between waves of the panel limit my ability to accurately measure the

effects of time-varying economic policies.

The low frequency of data collection employed in the LFS, restrict my analysis in several ways. Most importantly significant accuracy is lost in the measurement of unemployment duration due to clustering of responses. The loss in variation amongst respondents means that failure times are reported with some measurement error, as transitions between states are only recorded at the time of the next wave of the survey. All transition times in the sample are therefore recorded as later than the true transition time. This measurement-error varies within the sample in unknown ways. Additionally, there could be mis-measurement of unemployment duration amongst those making rapid transitions between states in the intervals between waves. This measurement-error rules out any accurate analysis of individual policies on the matching in the labour market.

Despite these shortcomings the LFS panel continues to offer the best available overview of the long-term dynamics of the job market in South Africa, over the period. In order to complete the analysis the data set was set up, using clear rules, to establish unemployment duration for all those who spent a portion of the sample unemployed.

As can be expected all individuals coded as either employed or economically inactive for the entire period under consideration were not included in the analysis. Additionally, individuals who did not display a transition to unemployment *and* weren´t continuously observed were excluded from the analysis. This latter exclusion is an attempt to minimize measurement error, as individuals with missing observations could potentially have transferred into the labour market during the missing period. However, if an individual transitioned before being censored, they are considered as unemployed up to the point of their transition.

2.3 Manually Censoring the Data

Table 2.8 Survival Data Before and After Censoring at 1300 Days of Analysis Time

	% censored		%change
	Before	**After**	
Rural-Urban			
rural	23.46	20.76	11.51
urban	20.7	18.24	11.88
Education			
No-High school	25.94	22.35	13.84
Some High School	19.08	16.51	13.47
Matric	19.26	17.43	9.50
NTC	24.57	24.36	0.85
Diploma/Certificate	21.86	19.93	8.83
Degree/PostGraduate Degree	30.94	30.14	2.59
Race			
African	19.94	17.06	14.44
Coloured	24.8	23.35	5.85
Indian/Asian	30.08	29.11	3.22
White	46.82	44.46	5.04
Gender			
Male	21.86	18.65	14.68
Female	21.69	19.7	9.17

Having established the "rules" of transition between states I can now turn to an overview the duration sample. For reasons elaborated upon in the following section of this paper, I will also censor all survival data at 1300 days of analysis time. Above a brief comparison of the survival data before and after censoring is given. As expected the amount of observed failures declines with censoring, while the number of observed individuals remains unchanged.

Effectively 833 of the 5039 observed transitions out of unemployment, or approximately 16.5%, have been censored and are included in the duration analysis only for the initial part of their unemployment spells. These observations therefore continue to contribute to the probability of remaining unemployed up to the date of censorship.

When manually censoring there arises the risk of biased estimation in several tests, including the rank based hypothesis tests used in section 3. Correcting this potential bias is made particularly difficult, as changes in censoring patterns are likely to be determined by the covariates. Individuals with lower education levels, for example, may be inherently more likely to have long periods of unemployment and therefore more likely to be censored. To account for this possibility both the Wald- and the log-

rank tests will be used, as the former is more sensitive to censoring and can therefore act as a robustness test (Cleves et al., 2004).

3 Evaluation of Unemployment Duration using Survival Analysis

The data assembled allows the use of duration models to analyse the interaction of several covariates with recorded unemployment duration. Before moving on to the analysis I will proceed by introducing the basic concepts of survival analysis as they relate to unemployment duration. While some knowledge of basic statistical concepts will be assumed, the reader need know little or nothing on the topic of survivor analysis. This section therefore gives a general overview of the analysis techniques used, while simultaneously applying these techniques to the data set assembled. Where appropriate interested readers will be referred to Statistical Appendix I, for a more in depth discussion of the technical points of survival analysis. Given that multiple statistical tests will be performed this section will focus primarily on giving an overview of trends in the data. A complete discussion of the results of all performed statistical tests can be found in Statistical Appendix II.

As a theoretical and technical basis this section relied heavily on the text by Cleves *et al.* (2004), which gives both a theoretical treatment of survival analysis and a comprehensive description of the implementation thereof in STATA. Additional theoretical insights into identification and specification of the Cox PH model were gained from van den Berg (2000). Readers who wish to gain more insights into the proportional hazards assumption of the Cox PH model are referred to Persson (2002). For a guide to the Aalen linear hazard model and the implementation thereof in STATA see Hosmer and Royston (2002).

3.1 Introduction to Survival Analysis

The application of survival analysis methodology to unemployment duration is not arbitrary, but instead, follows naturally from the search model framework discussed in Section 1.3. Theoretically predictions of the search model suggest that rate at which the unemployed workers find work should decrease over time, due to compositional effects and duration dependence. Compositional effect are the natural sorting that takes place as superior job candidates exit unemployment and remaining candidates have a decreased probability of receiving an offer of work. Duration dependence suggests that the long-term unemployment lose skills and decrease the intensity of their job search efforts over time (Blanchard & Diamond, 1990). Survival analysis, which explicitly models transition probabilities as a function of time spent in a given state is therefore preferable to other transition probability models, such as the probit or logit models, when assessing the empirics of the search model.

Conceptually survival analysis is primarily considered with the time spent in a given state of interest, such as unemployment, before transitioning to another. Transitions from the state of interest are referred to as failures. The term failure need not imply a negative outcome. In the case under

consideration, for example, an individual fails when transitioning from unemployment into employment. I focus my analysis on the variable T, a non-negative time variable measuring duration up to a failure event. An individual enters the sample under consideration at time t=0 when he either loses his job, or transitions into the labour force from the economically inactive population. I define cumulative density function F(t) and probability density function f(t) of T as:

$$F(t) = P[T < t]$$
$$f(t) = \frac{\partial F(t)}{\partial t}$$

More commonly in survival modeling the focus is on the survival function of t:

$$S(t) = P[T > t] = 1 - F(t)$$

Note that S(t) is defined as the probability that an individual has not failed by time t, or that he remains unemployed by for a time greater than t. Given the large proportion of discouraged work seekers in the South African economy it is to be expected that the estimated survival function will convergence to some stable level. Theoretically one can think of the level at which the survivor function converges as an equilibrium dictated by the search function.

3.1.1 The Kaplan- Meier Survivor Function Estimate

Estimation of the survival function will be done using the most standard non-parametric estimator, the Kaplan-Meier (1958) estimator. This estimator is calculated as:

$$\widehat{S(t)} = \prod_{j|t_j<t} \left(\frac{n_j - d_j}{n_j}\right)$$

where n_j is the number of individuals at risk at time t_j and d_j is the number of failures a time t_j[1].

Table 3.1 The Estimated Population Kaplan-Meier Survival Function

Time	Survivor Function	Std. Error	[95% Conf. Int.]	
1	0.9996	0.0002	0.999	0.9998
434	0.6466	0.0052	0.6364	0.6566
867	0.5485	0.0058	0.537	0.5598
1300	0.4768	0.0063	0.4645	0.489

[1] See Statistical Appendix I Note

Figure 3.1

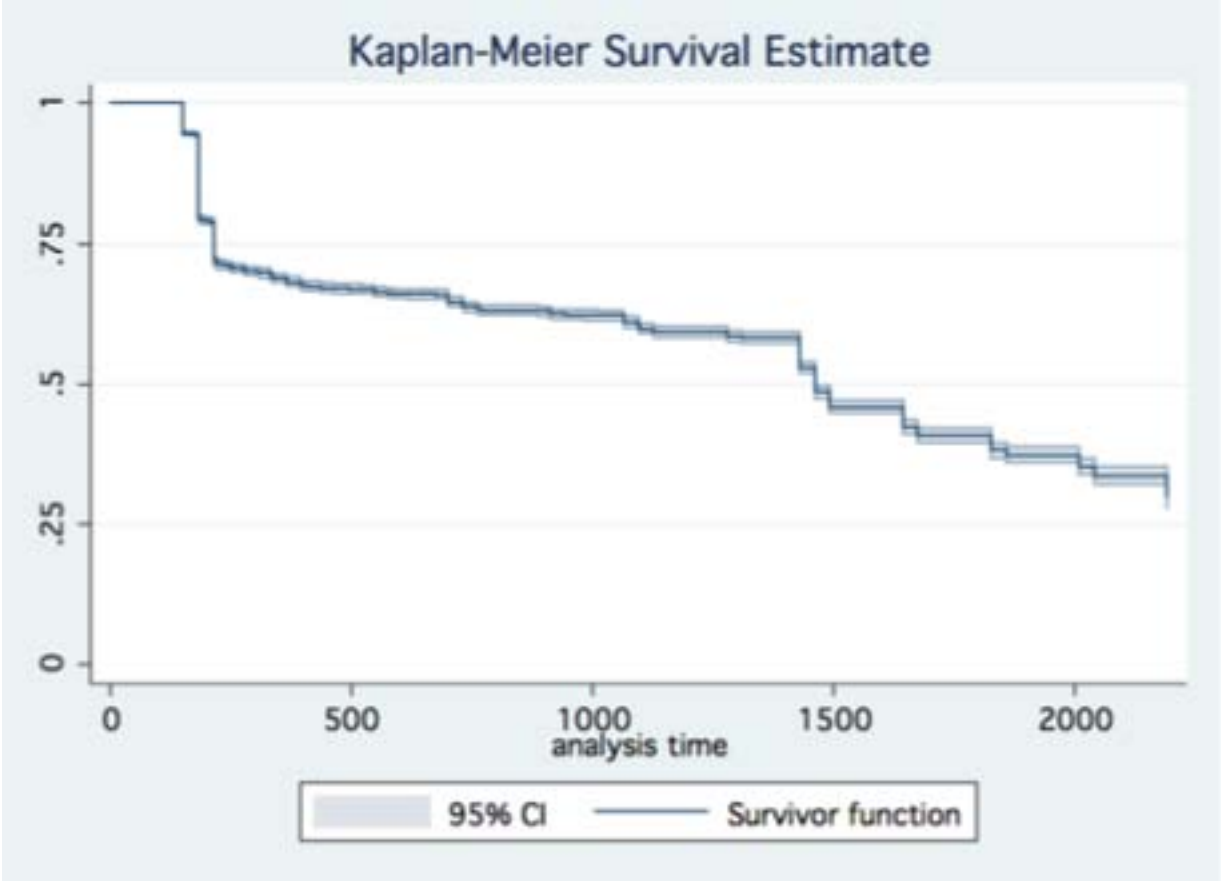


Table 3.1 and Figure 3.1 above give a graphical and descriptive summary of the Kaplan- Meier estimate for the survival function. A significant portion of the sample remains at risk for the minimal possible duration of 1 transition. Individuals who exit after one period can be thought of as frictionally unemployed, that is to say they remain classified as unemployed solely for a short period of time while looking for either their first job or an alternate source of employment.

The discontinuities in the graph above can be seen as the distances in days between waves of the survey. Surveys in this sample were not all equally spaced part. Note that the distribution seems to "smooth out" becoming less discontinuous for an extended period of the analysis time. This can be explained by the coding of the last worked variable which, where possible, relied on the respondents last noted date of employment to calculate the period unemployed. Such responses where coded by weeks, months or years depending on the length of unemployment. Notice the clear change in the slope of the Kaplan-Meier estimate around 1400 days of analysis time. I will return to the implications of this shift when investigating the hazard function in Section 3.1.2.

It can further be seen that the Kaplan-Meier estimate of the survival function suggests convergence to a long-term level of immobile unemployment. That is to say that some proportion of the sample

population do not leave unemployment for the duration of the sample. Extending table 3.I to multiple lengths suggests the survivor function stabilizes around 0.35. This implies that a startling 35% of the unemployed are predicted as not leaving the unemployed sample for the duration of the sample.

3.1.2 The Hazard Function

While the survivor function offers an accurate measurement of transition probabilities it leaves something to be desired in terms of intuitive ease of interpretation. For example the estimated survivor function given in Table 2.1 suggests that at analysis time of 831 days the probability of remaining unemployed is roughly 65%, however, this gives us no idea of the probability of transitioning back into employment.

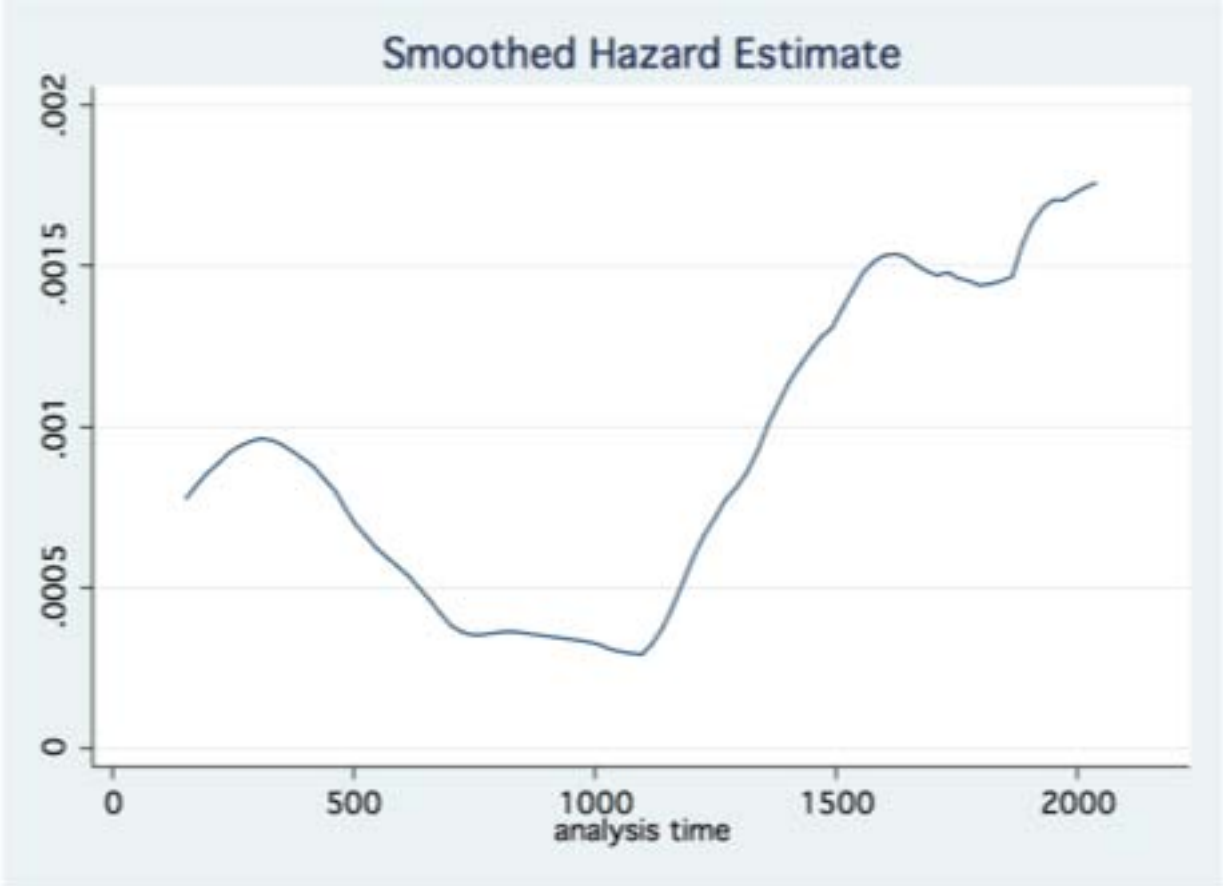
For this reason the remainder of this study will focus primarily on analyzing the estimated hazard function h(t), also known as the conditional failure rate. The hazard function is defined as the probability of failure at a given point in time conditional on survival up to that point. This definition can be stated mathematically as:

$$h(t) = \lim_{d \to 0} \Pr[t + d < T < t | T > t] = f(t)/S(t)$$

Theoretically the hazard function can be though of as the probability of transitioning into employment, or the level of labour absorption, at a given point in time. The emphasis given to the hazard rate in my analysis is not unique; the hazard rate forms the focal point of most econometric duration models. This follows naturally from economic theory which models economic decisions, at a given point in time, as conditional upon information received up to that point (van den Berg, 2000). Focusing my analysis on the hazard function therefore allows me to take into account changes in the economic environment over time, along with time varying covariates, which can affect behaviour in the job market[2].

[2] See Statistical Appendix I Note 2

Figure 3.2



* All estimated hazard functions are graphed using a Gaussian kernel density function.

At this point I will give my motivation for censoring the data. Recall that my *apriori* expectation was that the population hazard rate should be decreasing over time due to compositional effect and duration dependence. Figure 3.2 indicates that, contrary to the theory, the estimated hazard function displays a prominent U-shape over time. That is to say, individuals remaining in the labour force for an extended period face an increased probability of receiving an acceptable job offer and exiting the unemployed state.

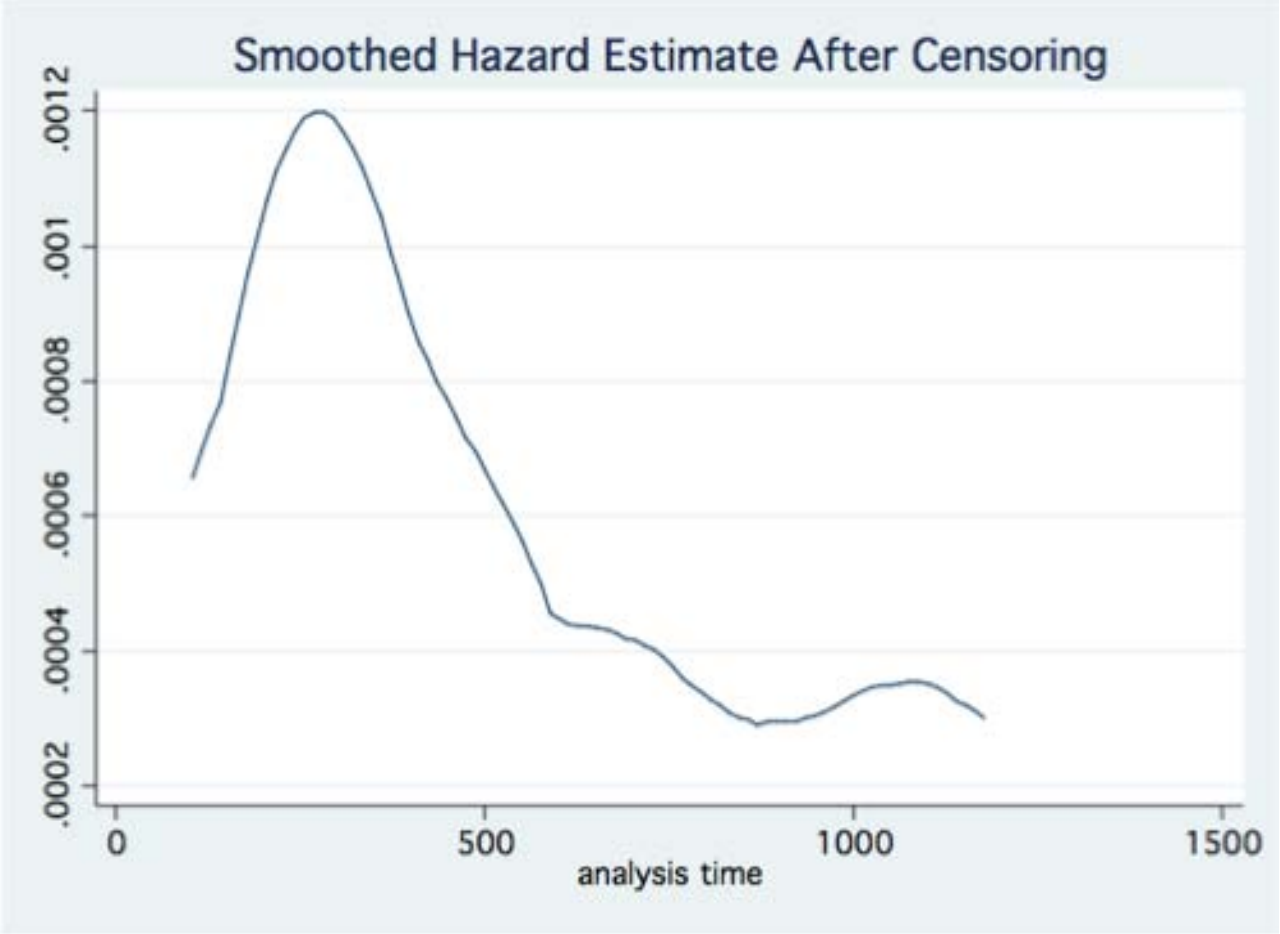
There are a number of potential explanations for the shape of the hazard function. The most feasible theoretical explanation would be to credit the behaviour of the hazard function to shifts in the reservation wage. This would imply that as unemployment becomes prolonged people lower the wages for which they are willing to work, thereby increasing both the probability of finding an acceptable job and the hazard rate. However, the reservation wage argument seems weak in light of the fact that the hazard rate only starts increasing after more than a thousand days of analysis time. One would reasonably expect that reservation wages should no longer be declining after such an extended period of unemployment. The behaviour of the survival function could also potentially attributed to measurement error caused by inaccurate responses to survey questions regarding unemployment

duration. However, this possibility has been discredited by a careful review of the data, which indicated that none of the individuals entering the data set after being unemployed for more than 3 years failed after more than 1000 days of observation.

Alternatively, wide scale censoring may bias the sample in some unknown way, leading to inaccurate measurements of the hazard functions amongst individuals surviving the longest. As noted earlier, the majority of the sample is followed for only two to three periods, far less than the period within which the hazard function starts to increase. A small proportion of failures at a late period can therefore act as an outlier in the relatively small subsample followed for more than 3 periods, exerting an upward bias on the estimated hazard rate.

The exact reason for which the hazard function displays this behaviour remains unclear. However, as these observations represent a relatively small section of the sample I will exclude the latest recorded failures by manually censoring the data. In this way I can continue to analyse unemployment duration within a limited period where the hazard function is more congruent with economic theory. For the purposes of analyzing unemployment duration the data was censored, limiting time observed after entering the unemployed state to 1300 days.

Figure 3.3



The estimated hazard function spikes around the 1st year of analysis time, equivalent to 1 wave in the survey. Unemployed individuals therefore had the highest probability of transferring into work in the period 0 to 6 months after the last recorded employment date. This initial spike in labour absorption is of special interest for my analysis, in Statistical Appendix 2 Section 3.2 I have decomposed survival estimates by individual covariates and indicated which subgroups had the initial upper hand in being re-employed.

3.1.3 The Nelson Aalen Cumulative Hazard Function Estimate

Similarly to the Kaplan-Meier estimate of the survival function I can non-parametrically estimate the cumulative hazard function using the Nelson (1972) and Aalen (1978) estimator[3]. This estimator is used in estimating the hazard function and will also be used later in the analysis for the estimation of the Aalen linear hazard model. The cumulative hazard function can be defined as:

$$H(t) = \int_o^t h(s)ds$$

[3] See Statistical Appendix I Note 3

Using the same definitions as above for t_j and d_j the Nelson-Aalen cumulative hazard estimator is defined as:

$$H(t) = \sum_{j|t_j<t} \frac{d_j}{n_j}$$

Table 3.2 The Estimated Nelson-Aalen Cumulative Hazard Function

Time	Std. Cumulative Hazard	Std. Error	[95%	Confidence Interval
1	0.0004	0.0002	0.0002	0.001
434	0.4263	0.0078	0.4113	0.4419
867	0.5897	0.0104	0.5696	0.6104
1300	0.7282	0.0129	0.7033	0.754

Figure 3.4

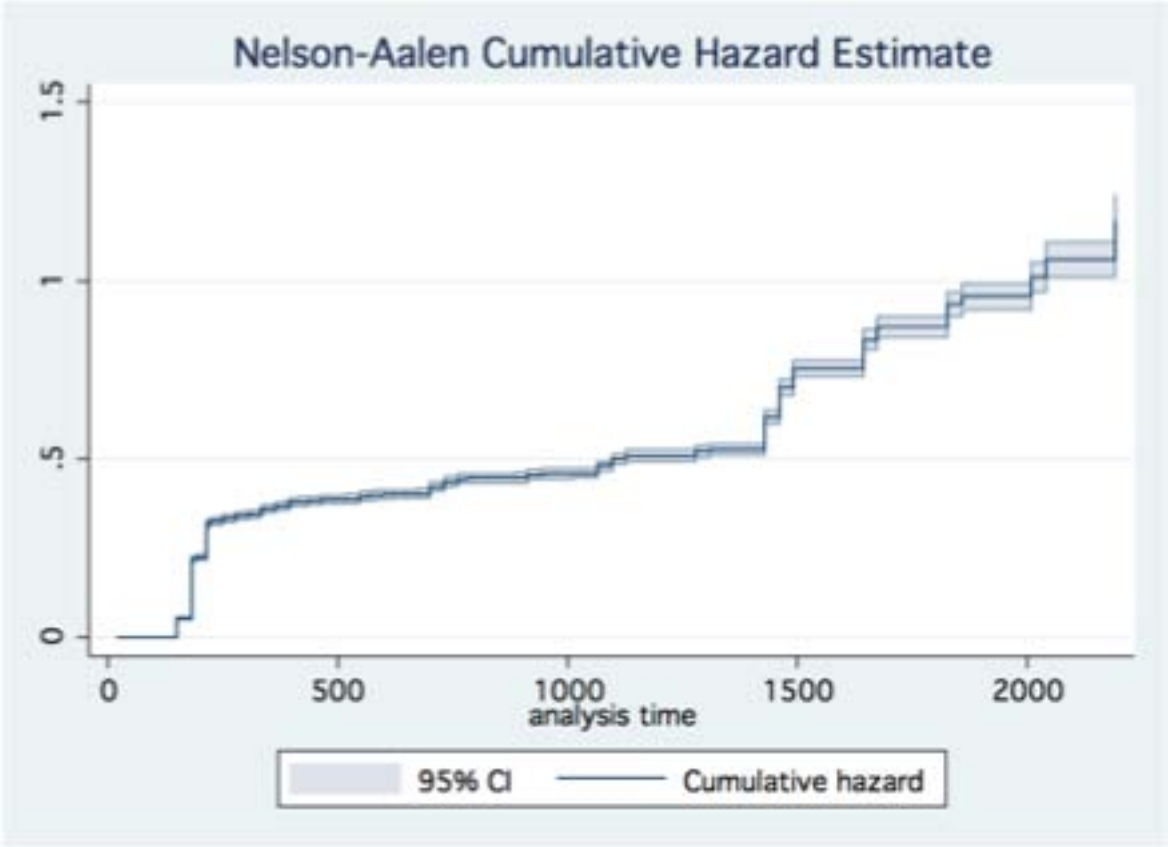


Table 3.2 and Figure 3.4 summarise the Nelson-Aalen estimate for the cumulative hazard function[4]. These estimates reinforce the findings of the Kaplan-Meier survival function estimate. As before, continuous periods in the data can be ascribed to the coding of the data.

[4] See Statistical Appendix I Note 4

3.2 Survival Analysis Based on Selected Covariates

Within the search framework I may expect that certain observable characteristic of an individual will affect her probability of receiving or accepting a job offer, and thereby exiting the unemployed population. Factors such as age, race, gender, education and the geographical distribution of labour have previously been linked to unemployment and will form the basis of my analysis (Banarjee *et al.* 2007).

In analyzing the differences in survival between subgroups I will follow several strategies. Firstly, a visual representation of the hazard function will be given. Secondly, a brief description of the estimated survivor and hazard functions will be used as a descriptive statistic of the data. Finally, I will perform hypothesis tests of equality of the estimators[5].

Before turning to descriptive analysis of survival and hazard rates between subgroups I will first establish that their survival rates do in fact differ. There exist several non-parametric tests, which test the global equality of the survivor function. That is to say the hazard of failure is equally distributed over the entire duration of the sample. These tests differ primarily in respect of their treatment of the time-based weighting they use to compare the two survival functions.

I will limit my analysis to the Log-Rank and the Wilcoxon tests, both of which are rank tests comparing survival times. The Log-Rank test equally weighs all observations while the Wilcoxon test gives larger weightings to observations which fail earlier. The rank tests can also be stratified to control for differences in a second variable. In this way I can continue to control for differences based on one indicator variable, such as gender, while controlling for differences in another, such as race.

Performing both the log-rank and Wilcoxon test the hypothesis of equality amongst the survivor based on differences in gender, education, race, education or rural-urban status is rejected. This result is also robust to stratification based on the other covariates. These tests therefore confirm statistically significant differences in survival rates for the covariates under consideration[6].

The estimates of the survival function furthermore suggest convergences in the survival rates as decomposed by gender and urban rural differences. Conversely the educational, racial, provincial and youth/non-youth decomposition of the survival function suggest that the long-term survival rates diverge when decomposed by these categorical variables. The rankings of survival estimates, which

[5] For an extended analysis see Statistical Appendix II Section 3.2
[6] See Statistical Appendix I Note 5

can be thought of as an ordering of labour market conditions by covariate, are largely congruent with the findings of other studies such a that of Fourie(2011). One notable exception is that the survival rates of youth converge to a lower level than that of non-youths suggesting that, in the medium- to long-term, youths are more likely to exit the labour force than non-youths.

3.3 Modeling the Hazard Function

Having completed my initial investigation into a descriptive analysis of the data I now turn my attention to modeling a hazard function for the sample. Such an exercise will allow me to model hypothetical individuals in the sample and compare labour market experience.

Duration models can generally follow either parametric or non-parametric estimation strategies. Unfortunately the large lags between survey responses means that my data is somewhat imprecise. Making the distributional assumptions, necessary for the use of a parametric model, is therefore somewhat difficult. I will limit my analysis to the non-parametric Cox Proportional Hazard model.

3.3.1 The Cox Proportional Hazard Model: An Introduction

The Cox (1972) Proportional Hazard model remains the most popular method of analyzing manner in which observed variables alter the hazard rate due to it elegance and computational feasibility (Cleves *et al.* 2004). The model assumes that covariates are associated with a multiplicative shift of the baseline hazard function. This implies that the conditional hazard rate for individual j is:

$$h(t|\boldsymbol{x_j}) = h_0(t)\exp\left(\boldsymbol{x_j\beta_x} + (t\boldsymbol{x_j})\boldsymbol{\beta_{xt}}\right)$$

Note that in the above equation the conditional hazard function is determined by an underlying baseline hazard, which is multiplicatively altered by the values of the covariates. Importantly the Cox model makes no assumption on the functional characteristics of the baseline hazard function. Additionally tied failures, where multiple failures were recorded at the same time, were handled according to the Breslow (1974) method[7]. During the Cox PH analysis that follows I will be interested primarily in accurately comparing hazard rates between subgroups based on multiple covariates.

The Cox-PH model offers several possibilities for my current analysis. Firstly, I can us it as a simple descriptive statistic to differentiate between alternate groups, as defined by a single covariate. Such an analysis, while limited, allows for a simplistic comparison between groups and will form the departure point of my analysis. Additionally, one may wish to ascertain whether differences in hazard ratio based on one covariate, such as race, persist once I have accounted for another, such as level of education. In this way, I can establish the persistence and magnitude of effects such as racial

[7] See Statistical Appendix Note 6

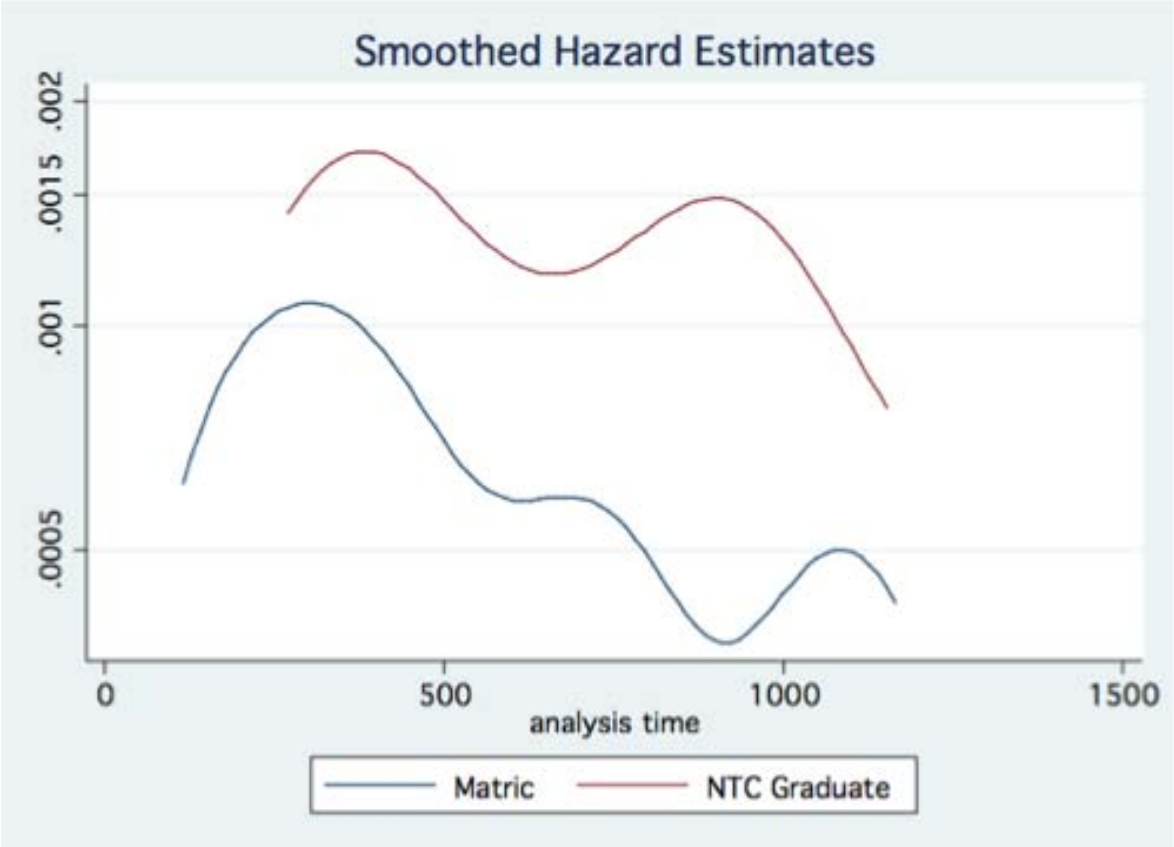
discrimination in the labour market while simultaneously ascertaining which demographic group experiences the worst, and the best, job prospects.

Importantly the categorical nature of most of the demographic variables I use requires the establishment of a baseline for comparison. Baselines in my case were not chosen haphazardly but given primarily to outlying subgroups. These subgroups were identified graphically using the smoothed hazard estimates given in section 3.2 of Statistical Appendix II. Note that the non-linear relationship between the hazard rate and unemployment means the fourth quintile is used as the baseline .

3.3.2 The Proportional Hazard Assumption

As has been mentioned the assumption of multiplicative shifts in the conditional hazard rate is crucial for the use of the Cox PH model. As such I should first attempt to estimate the viability of the model for the sample. Where the proportional hazard assumption is not met, the Cox PH model no longer accurately measures differences in hazard rates over the entire sample. Instead, estimates of the hazard ratio will be similar to the exact calculation of the geometric average hazard ratio (Persson, 2002).[8]

Figure 3.5



[8] See Statistical Appendix I Note 7

Under the proportional hazards assumptions the effects of covariates should be associated to a multiplicative shift in the hazard function and should therefore be time invariant. Estimates should behave in similar fashion, forming two lines of similar shape which do not cross. Divergence, or convergence, of two hazard functions disqualify a subsample from analysis. As an example, consider the clear divergence in estimated hazard functions for NTC graduates and matrics, given in figure 3.5. Additionally the estimated hazard functions cannot cross, as this would be a violation of the proportional hazard assumption.

The importance of evaluating the proportional hazard assumption extends beyond assessing the applicability of the Cox PH model. Determining the behaviour of the hazard ratio between subgroups over time allows us to analyse whether labour market characteristics faced by different subgroups differ, not only on aggregate, but dynamically over time.

Two methodologies exist for testing the proportional hazards assumption; numerical tests or graphical analysis. Homer and Lemeshow (1999), recommend the use of numerical tests to avoid the subjectivity that can influence the conclusions of a graphical analysis. Others reject numerical tests in favour of graphical procedures, arguing that for a large enough data set all numerical tests will reject the proportionality assumption for even the smallest deviations from proportionality (Klein and Moeschberger, 2003). Regardless of the methodology used, testing the proportional hazard assumption is complicated significantly by measurement error in my data set. Measurement error reduces the power of all numerical tests in detecting non-proportionality, and significantly increases the difficulty of graphically assessing proportionality (Persson, 2002).

Graphical tests of the proportional hazards assumption

Cleves *et al.* (2002) suggest two methods of testing the proportional hazard assumption graphically. The first uses a transformation of the of the Kaplan-Meier estimated survivor function while the second compares the Kaplan-Meier estimates to the Cox PH models estimated survival function.

The first graphical test makes use of the fact that, under the assumptions of the Cox PH model, the Survival function is defined as:

$$S(t|x) = S_0(t)\exp(\mathbf{x\beta_x})$$

To plot the transformation

$$-\ln[-\ln\{S(t|x)\}] = -\ln[-\ln\{S(t|x)\} - \mathbf{x\beta}_x]$$

Under the proportional hazards assumption a graph of this function by subgroups should yield parallel lines.

The second test compares the Kaplan-Meier estimates of the survival function to those of the Cox PH model. The Kaplan-Meier imposes no parametric assumptions on the survival function while the Cox PH model imposes the assumption of proportionality. Large differences between these estimates are therefore indicative of non-proportionality.

Figure 3.6

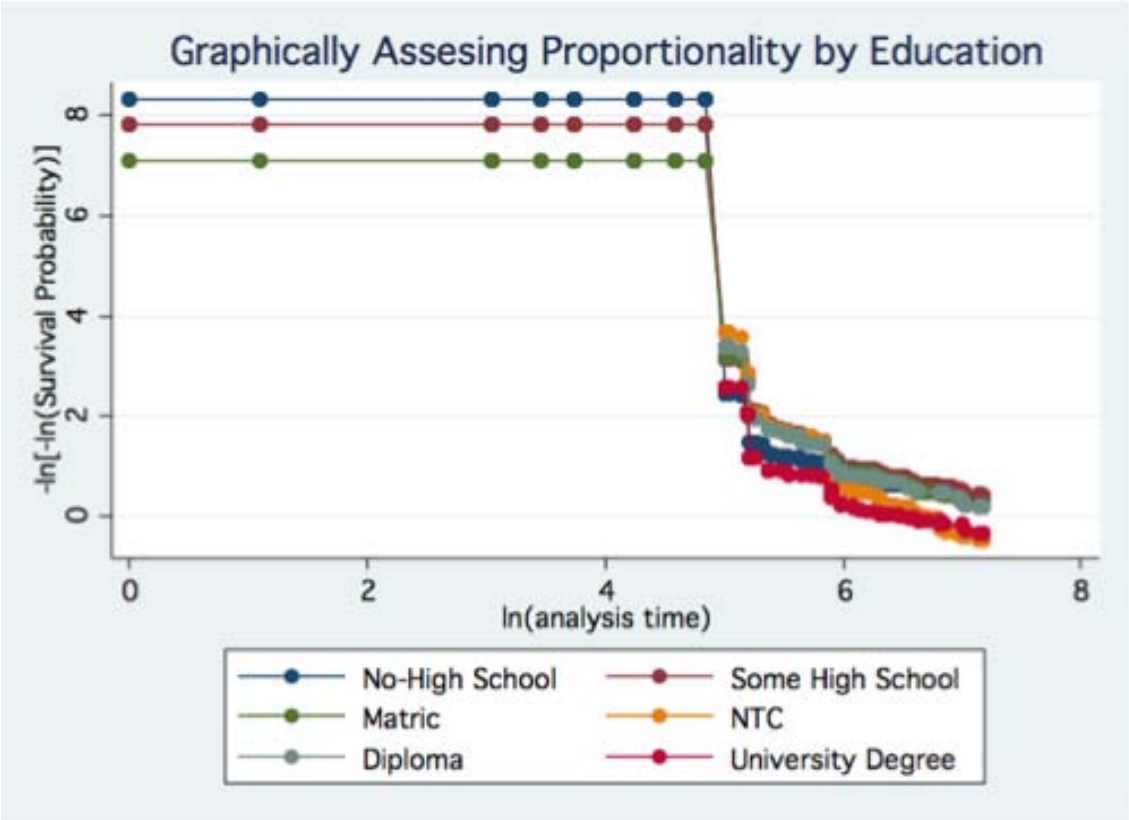
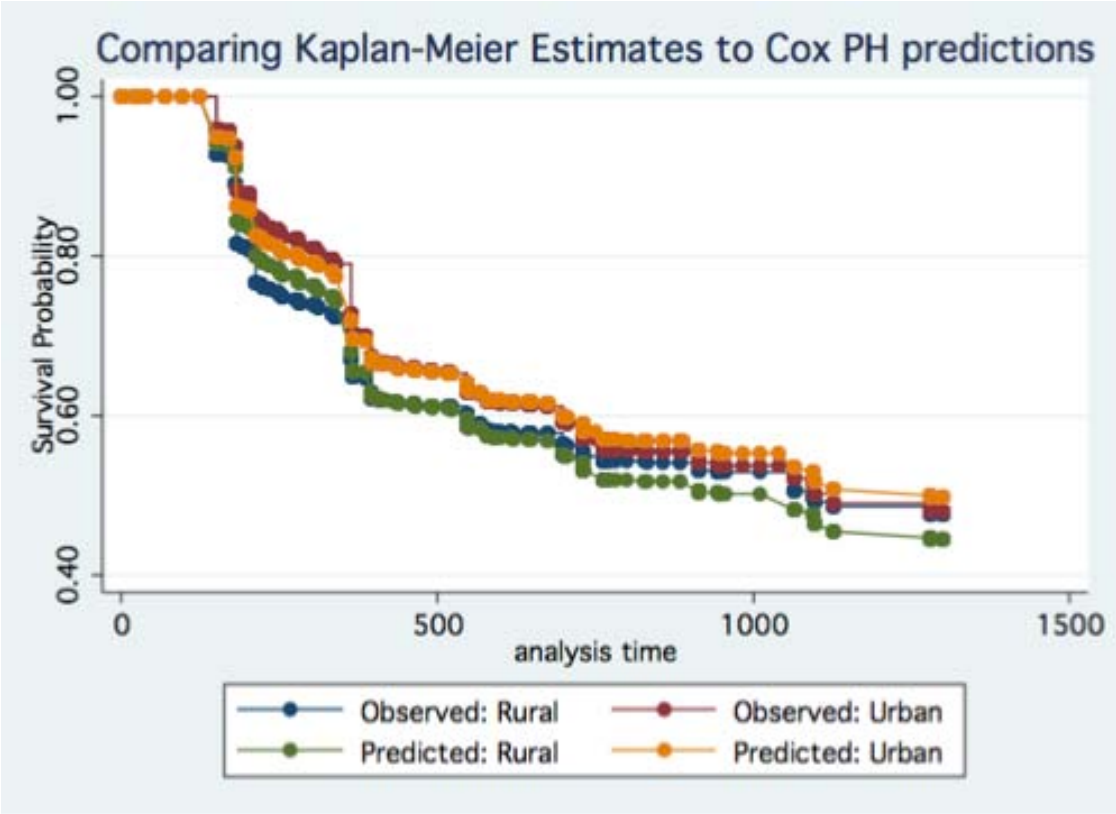


Figure 3.7



As stated before, and illustrated in the above figures the graphical analysis of the proportional hazards assumption, based on the methodologies listed above, is rendered difficult by measurement error and the multitude of categorization of the multinomial variables.

A simplistic alternative graphical test of the proportional hazard assumption is by visual inspection of a logarithmic transformation of the estimated hazard functions (Persson, 2002). These figures largely confirm the findings of section 2.2. I conclude that for subgroups where hazard rates cross, such as gender or the urban indicator, accurately modeling the hazard rates for different subsets of the population becomes unfeasible using the Cox model. Note however that these are smoothed estimates of the hazard function. Any inference based on these figures are therefore tentative at best.

Numerical tests of the proportional hazards assumption

While several numerical tests exist to test the proportional hazards assumption I will limit my analysis to the time-dependent covariate test of Cox (1972) and the weighted Schoenfeld residual score test proposed by Grambsch and Therneau (1994). Simulations have shown these test to have high power in detecting non-proportionality in a variety of cases (Persson, 2002).

The Cox (1972) time dependent covariate test remains the most simple way of directly testing the proportional hazard assumption. This test includes a time interaction term into the original Cox PH

model such that it becomes:

$$h(t|\boldsymbol{x_j}) = h_0(t)\exp\left(\boldsymbol{x_j\beta_x} + (t\boldsymbol{x_j})\boldsymbol{\beta_{xt}}\right)$$

Where the coefficient of the interaction term is significant a linear relationship exists between time and the hazard ratio and the proportional hazard assumption is violated. I can therefore simply complete a hypothesis test of coefficient of the time interaction variable above differs significantly form zero to establish whether or not proportionality holds.

An alternate test of the proportionality assumption is the Grambsch and Therneau(1994) test based on the Schoenfeld (1982) residuals of the fitted Cox model. The Schoenfeld residuals can be thought of as the difference between the expected and the observed times of failure as predicted by the Cox-PH model. Where these residuals display a relationship with time I can conclude that the relationship between the hazard rate and the covariate varies over time, and the proportionality assumption does not hold. The measurement error in my data is likely to adversely effect estimates of the Schoenfeld residuals. Results of the Grambsch and Therney test therefore need to be treated with some caution and are given primarily as a robustness check of the Cox proportionality test results[9].

[9] For an extended analysis see Statistical Appendix II Section 3.3.3

3.3.3 Cox PH Hazard Ratio Estimates and Diagnostic Test Results

Table 3.3.1 Cox PH Model Estimates with Diagnostic Tests by Individual Covariate

	Cox PH Model Hazard Ratio estimate			**Cox diagnostic test**		**Grambsch-Therneau diagnostic test**		
	Hazard ratio	Std Error	p>|z|	Beta xt	p>|z|	chi sq	0.000	DF
Male-Female	1.082	0.034	0.011	-0.00037	0.002	9.79	0.002	1
Education						161.68	0.000	5
No-high school	0.605	0.061	0.000	0.00108	0.000	5.98	0.015	1
Some high school	0.477	0.048	0.000	0.00157	0.000	0.04	0.835	1
Matric	0.538	0.056	0.000	0.00275	0.000	1.97	0.161	1
NTC	0.890	0.103	0.311	0.00154	0.000	14.74	0.000	1
Diploma	0.561	0.075	0.000	0.00102	0.024	1.3	0.255	1
Rural-Urban	0.859	0.027	0.000	0.00089	0.000	48.99	0.000	1
Provincial						81.11	0.000	8
Eastern Cape	0.770	0.042	0.000	-0.00189	0.000	53.83	0.000	1
Northern Cape	0.587	0.047	0.000	-0.00028	0.347	0.48	0.487	1
Free State	0.534	0.035	0.000	-0.00029	0.215	0.67	0.414	1
KwaZulu Natal	0.701	0.039	0.000	-0.00059	0.007	6.07	0.414	1
North West	0.468	0.030	0.000	-0.00050	0.040	3.14	0.076	1
Gauteng	0.537	0.032	0.000	-0.00004	0.862	0.18	0.674	1
Mpumalanga	0.634	0.039	0.000	-0.00077	0.001	9.27	0.002	1
Limpopo	0.492	0.032	0.000	-0.00016	0.512	0.11	0.737	1
Racial Groupings						16.02	0.001	3
Asian	2.026	0.177	0.000	0.00050	0.001	8.74	0.003	1
Coloured	1.561	0.065	0.000	0.00011	0.785	0.07	0.790	1
White	2.982	0.167	0.000	-0.00077	0.008	4.98	0.256	1
Age based differences						103.75	0.000	3
Age	0.949	0.007	0.000	-0.00002	0.921	16.46	0.000	1
Age^2	1.001	0.000	0.000	-0.00010	0.040	23.28	0.000	1
Youth	1.400	0.089	0.000	0.00000	0.437	0.3	0.585	1

Table 3.3.2 Cox PH Model Estimates with Diagnostic Tests Controlling for Covariates

	Cox PH Model Hazard Ratio estimate			**Cox diagnostic test**		**Grambsch and Therneau diagnostic test**		
	Hazard ratio	Std Error	p>|z|	Beta xt	p>|z|	chi sq	0.000	DF
Global test						318.51	0.000	25
Female-Male	1.169	0.037	0.000	-0.00049	0.000	10.75	0.001	1
Education								
No-high school	0.701	0.076	0.001	-0.00130	0.003	8.06	0.005	1
Some high school	0.635	0.067	0.000	-0.00088	0.042	2.83	0.092	1
Matric	0.700	0.075	0.001	-0.00030	0.494	0.05	0.827	1
NTC	1.132	0.134	0.297	0.00073	0.115	4.25	0.039	1
Diploma	0.657	0.089	0.002	0.00004	0.932	0.55	0.459	1
Rural-Urban	0.720	0.027	0.000	0.00062	0.000	18.35	0.000	1
Provincial								
Eastern Cape	0.619	0.066	0.000	-0.00201	0.000	13.19	0.000	1
Northern Cape	0.336	0.034	0.000	-0.00027	0.487	0.02	0.884	1
Free State	0.367	0.036	0.000	0.00040	0.242	3.05	0.081	1
KwaZulu Natal	0.647	0.071	0.000	-0.00042	0.289	0.57	0.452	1
North West	0.539	0.062	0.000	-0.00018	0.668	0.06	0.814	1
Gauteng	0.575	0.045	0.000	0.00016	0.560	0.96	0.326	1
Mpumalanga	0.337	0.034	0.000	-0.00016	0.665	0.02	0.883	1
Limpopo	0.334	0.045	0.000	-0.00130	0.020	2.71	0.100	1
Racial								
Groupings								
Indian	1.434	0.084	0.000	0.00010	0.657	0.18	0.673	1
Coloured	2.200	0.209	0.000	-0.00064	0.156	6.45	0.011	1
White	2.498	0.163	0.000	-0.00162	0.000	37.73	0.000	1
Age based differences								
Age	0.956	0.007	0.000	-0.00002	0.718	9.75	0.002	1
Age^2	1.001	0.000	0.000	0.00000	0.596	7.85	0.005	1
NonYouth- Youth	1.309	0.085	0.000	0.00006	0.804	0.52	0.472	1

Our Cox PH analysis of employment is similar to other analysis of demographic determinants of unemployment such as that of Bhorat and Oosthuizen (2005). My analysis has additionally included local unemployment rates and a youth indicator variable. The former was estimated using quintiles of unemployment for the entire sample while the later was an indicator variable taking on a value of 1 for all individuals who remained below 30 years of age for all recorded observations.

I begin my analysis with covariates taken in isolation. These estimates can be considered as a simple measurement of differences between demographic subgroups, as defined by a single covariate. As expected statistical differences exist in the estimated hazard ratios between subgroups for all the variables considered. Next I expand my analysis to include all covariates into a single estimation. These measurements can be thought of as a more accurate measure of the contribution of each covariate to the equilibrium of the search function.

Estimation results are not given in terms of the beta coefficients of the model given in section 3.2.1. Instead, I will focus on the hazard ratio, or the relative hazard of a individual with a covariate value of x+1 relative to that of an individual with a covariate value of x. In the case of binomial covariates the hazard rate can be simplified as:

$$HR = \frac{h(t|x = 1)}{h(t|x = 0)}$$

A individual with a covariate value of 1 will therefore have a hazard rate equal to HRx100% the hazard rate of one with a value of 0. As an example consider the estimated HR for the gender indicator variable in Table 3.3.1 above, I can conclude from this estimate that amongst women the hazard of exiting unemployment was on aggregate 1.082*(100%)=108.2% that faced by men. Women therefore had an 8.2% higher probability being matched to a job on aggregate.

By combining the results of the Cox PH model with those of the Cox diagnostic test I can also gain some insight into the behaviour of the hazard ratio over time. Consider the case where a covariate is associated with an increase (decrease) in the hazard rate the Cox PH model will estimate a hazard rate greater (less) than one. In this case a positive coefficient estimate for the time interaction term in the Cox diagnostic test is indicative of divergence (convergence) of hazard rates over time. Similarly, a negative coefficient estimate for the time interaction term is associated with convergence (divergence) of the hazard term.

Table 3.4 Interpreting the Cox PH Model Diagnostics

Hazard Ratio	Interaction term coefficient	Relationship of the hazard ratio
HR>1	Beta<0	Convergence
	Beta>0	Divergence
	Beta=0	Proportionality
HR<1	Beta<0	Divergence
	Beta>0	Convergence
	Beta=0	Proportionality

To clarify interpretation consider once again the estimates given in table 3.3.1 for the gender indicator variable. As noted the estimated hazard ratio is 1.082>1, while the coefficient for the time interaction term is -0.00037<0. I can therefore conclude that the hazard rates, as decomposed by gender, are converging over time. A summary of the rules of interpretation is given in Table 3.4 above.

The results of the Cox PH test and associated diagnostic test that the proportionality assumption is rejected for all categorical variables. The amount of individual covariates testing positive for non-proportionality is reduced, but not eliminated, when controlling for other covariates. The HR estimates of the Cox model itself yield little surprises apart from the fact that NTC graduates do not statistically differ from university graduates and that the relationship between age and the hazard rate is convex.

3.4 The Aalen Linear Hazard Model

The residual analysis performed on the Cox PH model gave insights into the existence, and aggregate movement, of time varying effects of the hazard ratio. However, the analysis performed remains somewhat unsatisfactory from an analytical point of view. Specifically, I may wish to extend my analysis to establish the non-monotonous relationships in the estimated hazard rates. To this purpose I will make use of an alternate non-parametric model; Aalen´s (1989) proportional hazard regression model (Hosmer & Royston, 2002)[10].

While, similarly to the Cox-PH model, the Aalen linear hazard model makes no assumptions on the baseline hazard it differs with respect to the assumed functional form of the hazard function. The Aalen

[10] See Statistical Appendix I Note 8

linear hazard model assumes that the hazard function is of form:

$$h(t|\boldsymbol{x_j}) = h_0(t) + \boldsymbol{x_j}\boldsymbol{\beta_x}(t)$$

Note that in the above example the estimated coefficients are allowed to vary over time, additionally differences in hazard ratios by covariate at a given time are linearly additive.

When analyzing the relationship of the hazard ratio using the Aalen linear hazard model I will not focus on a single estimate of the coefficients, instead I am interested in dynamic changes over time. For this reason I will focus only on the graphical output of the Aalen model to ascertain whether relative hazards converge, diverge or stabilize at some level. Once again my analysis will focus on both relative hazards as measured by a single covariate and relative hazards where controlling for all covariates, as in table 3.3.2. All graphs of the coefficient estimates include both an estimate of the coefficient for the covariate of interest, a 95% confidence interval thereof and a red line at a coefficient value of zero. Where the confidence interval crosses the red line we can conclude that the coefficient estimate does not differ significantly form zero.

3.4.1 Gender Based Differences

Figure 3.8.1.a No Control Variables

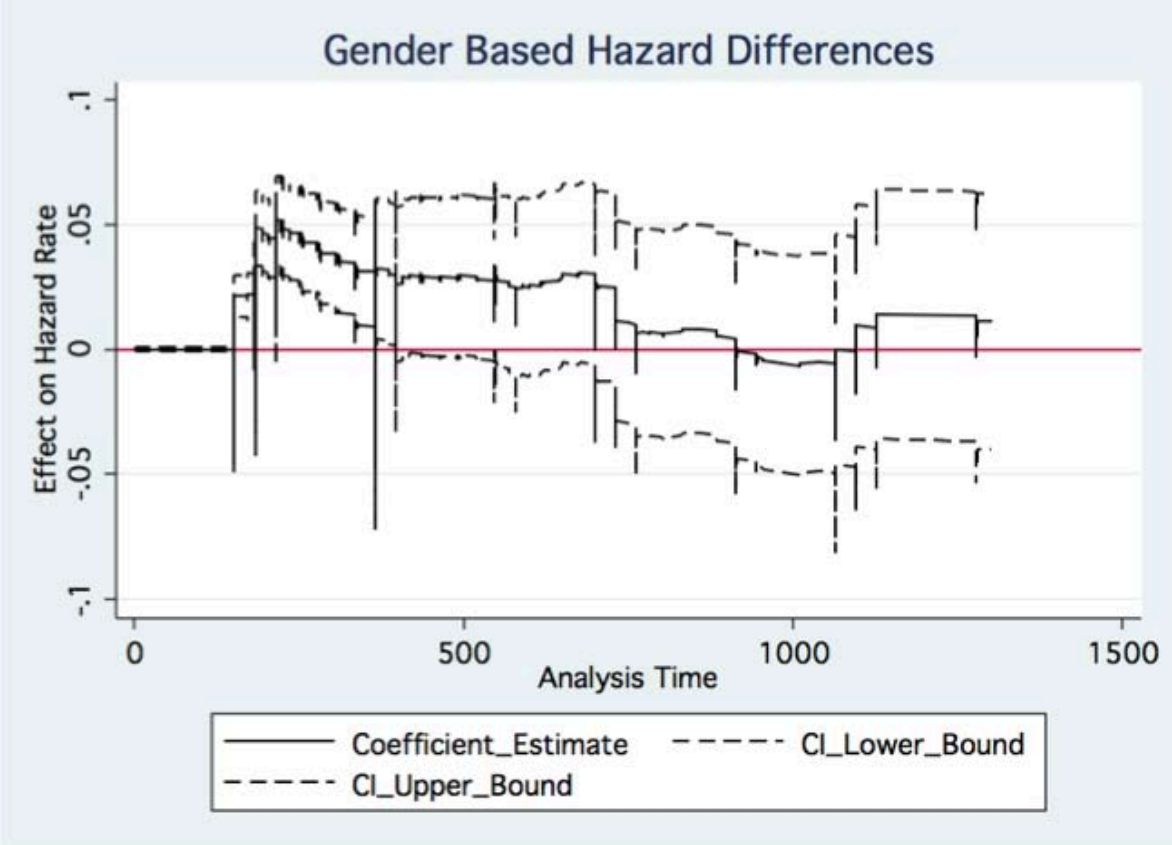
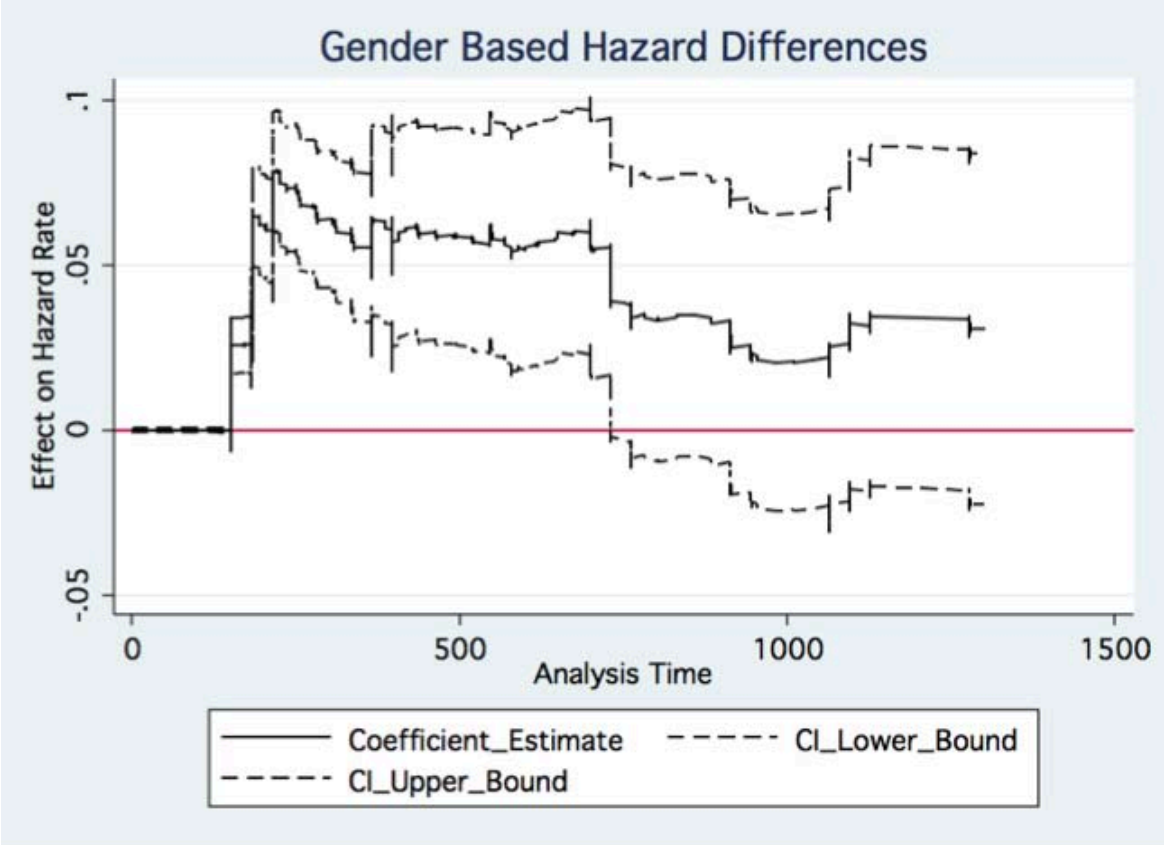


Figure 3.8.1b With Control Variables



As with the Cox PH model the Aalen linear hazard model finds only a small difference in the hazard rate between men and women. Additionally this difference tends to zero over time suggesting that the hazard rates of men and women are essentially identical in the long-run. Once I control for the other covariates the magnitude of the gender effect increases considerably but continues to converge to a non-statistically significant level over time.

3.4.2 Education Based Differences

Figure 3.8.2.1.a No Control Variables

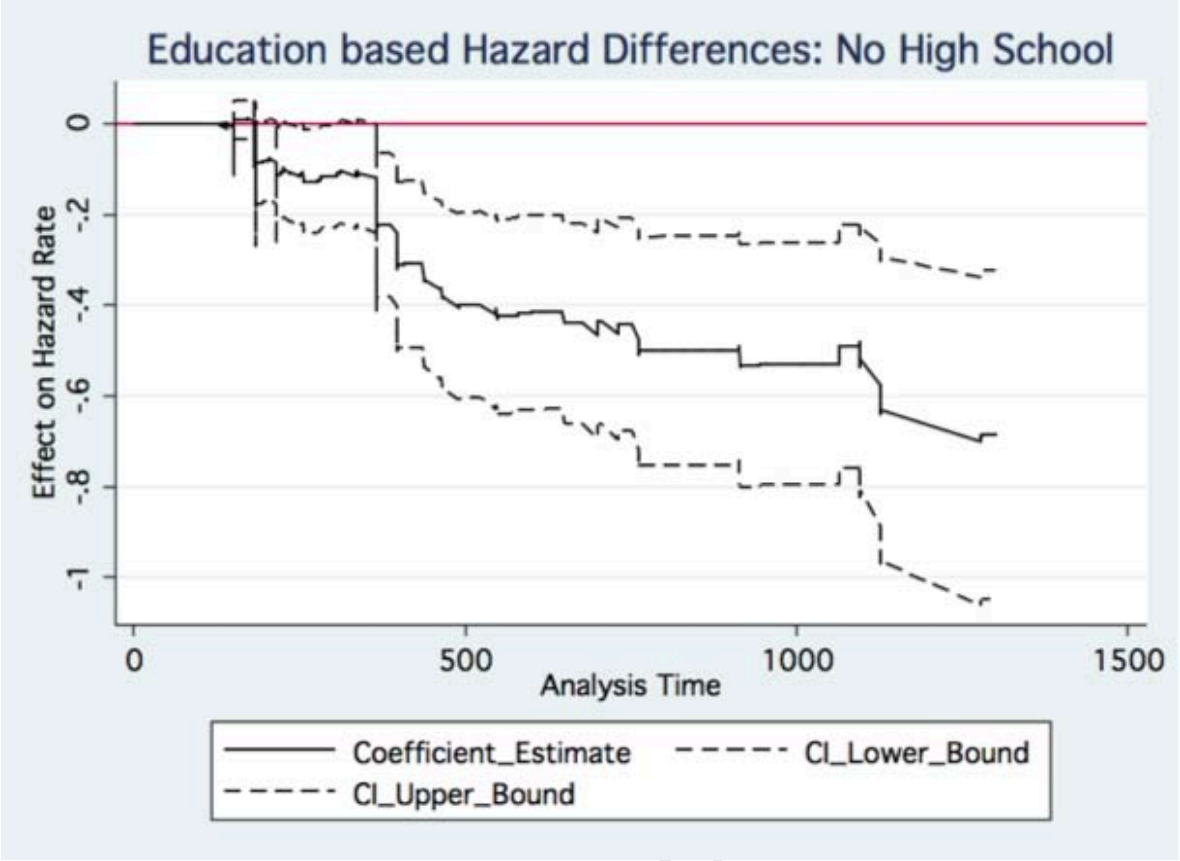


Figure 3.8.2.1b With Control Variables

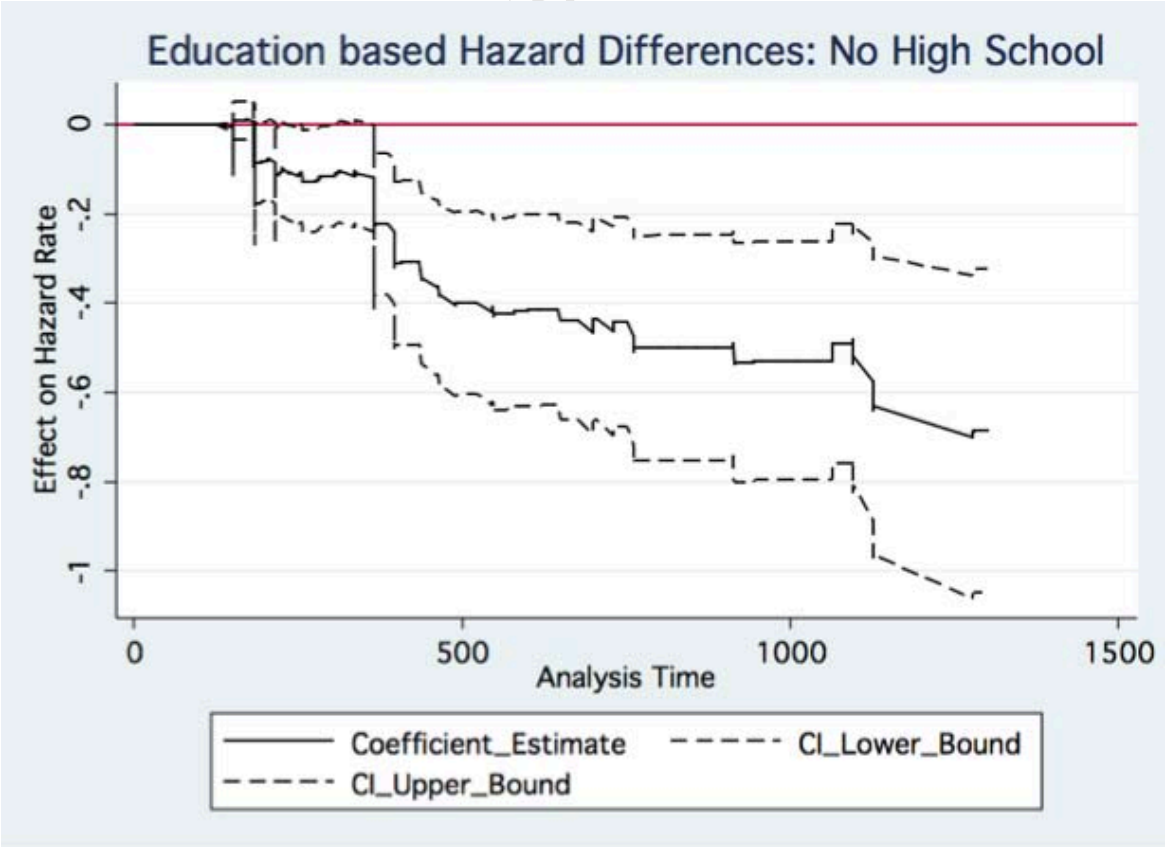


Figure 3.8.2.2.a No Control Variables

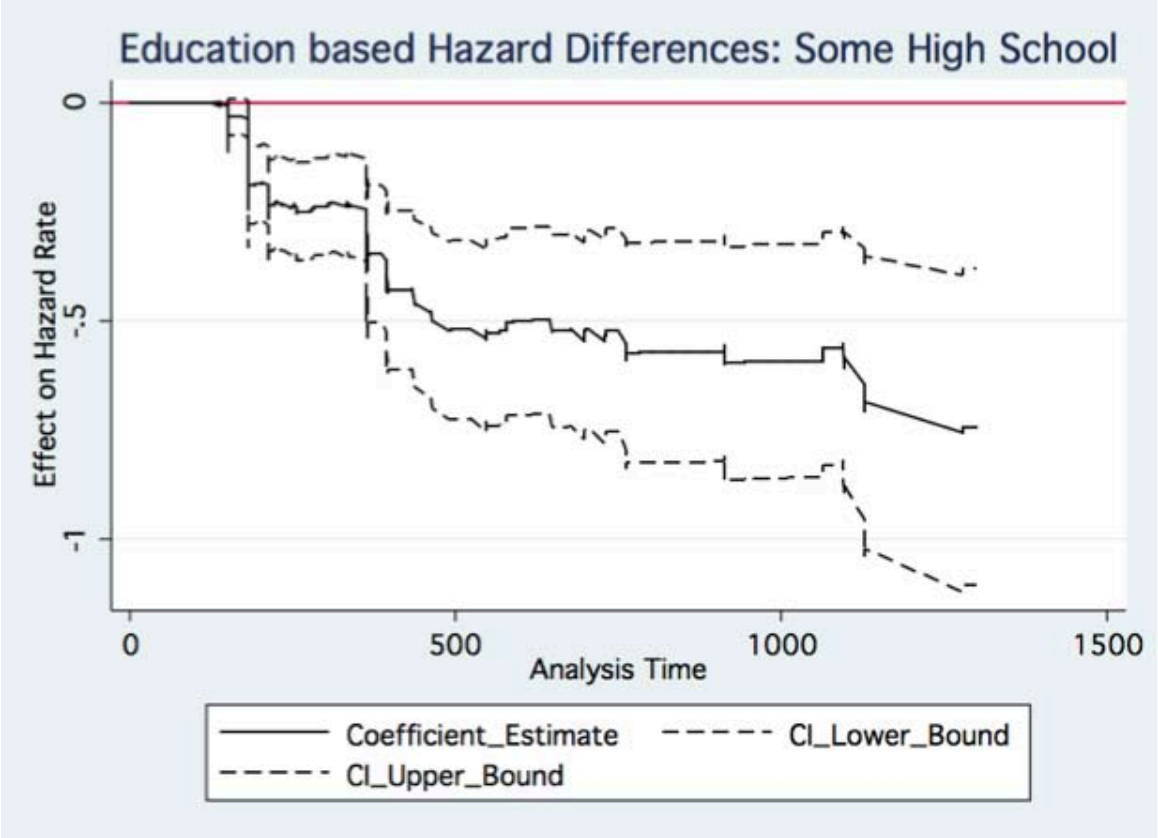


Figure 3.8.2.2b With Control Variables

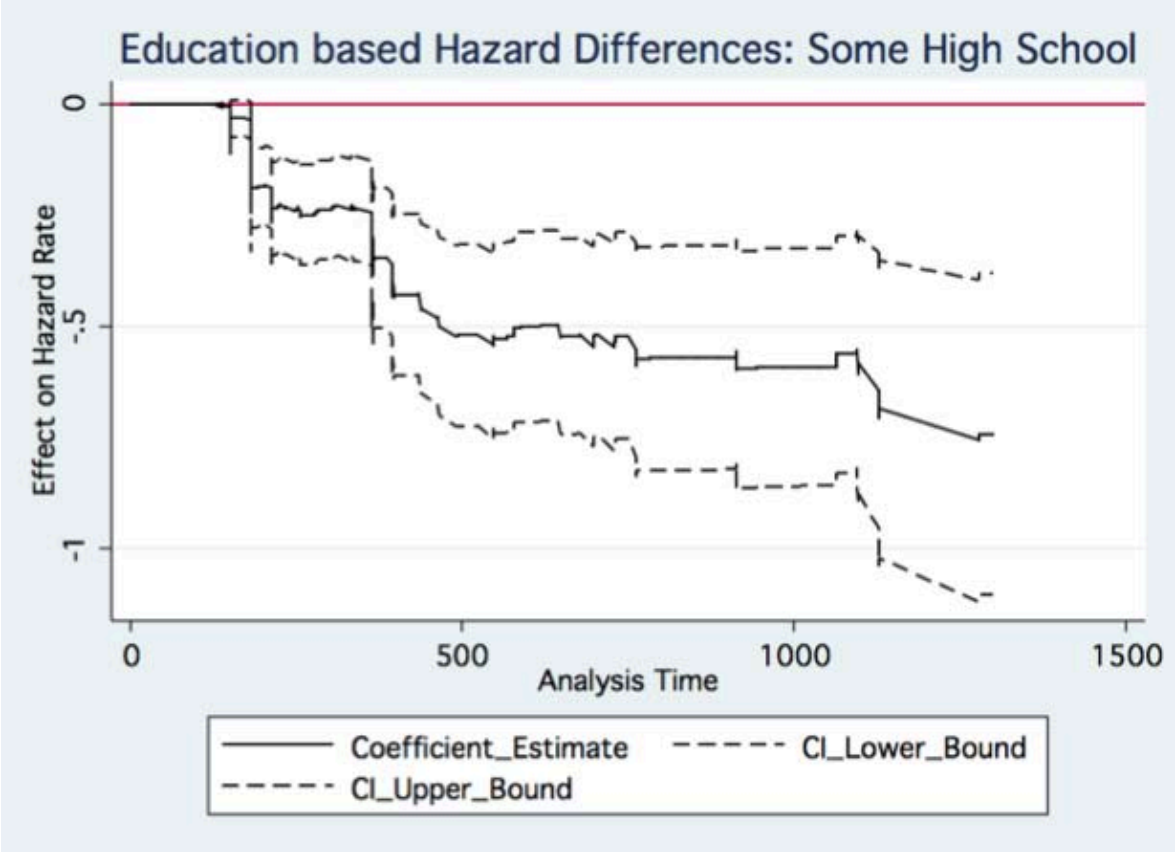


Figure 3.8.2.3.a No Control Variables

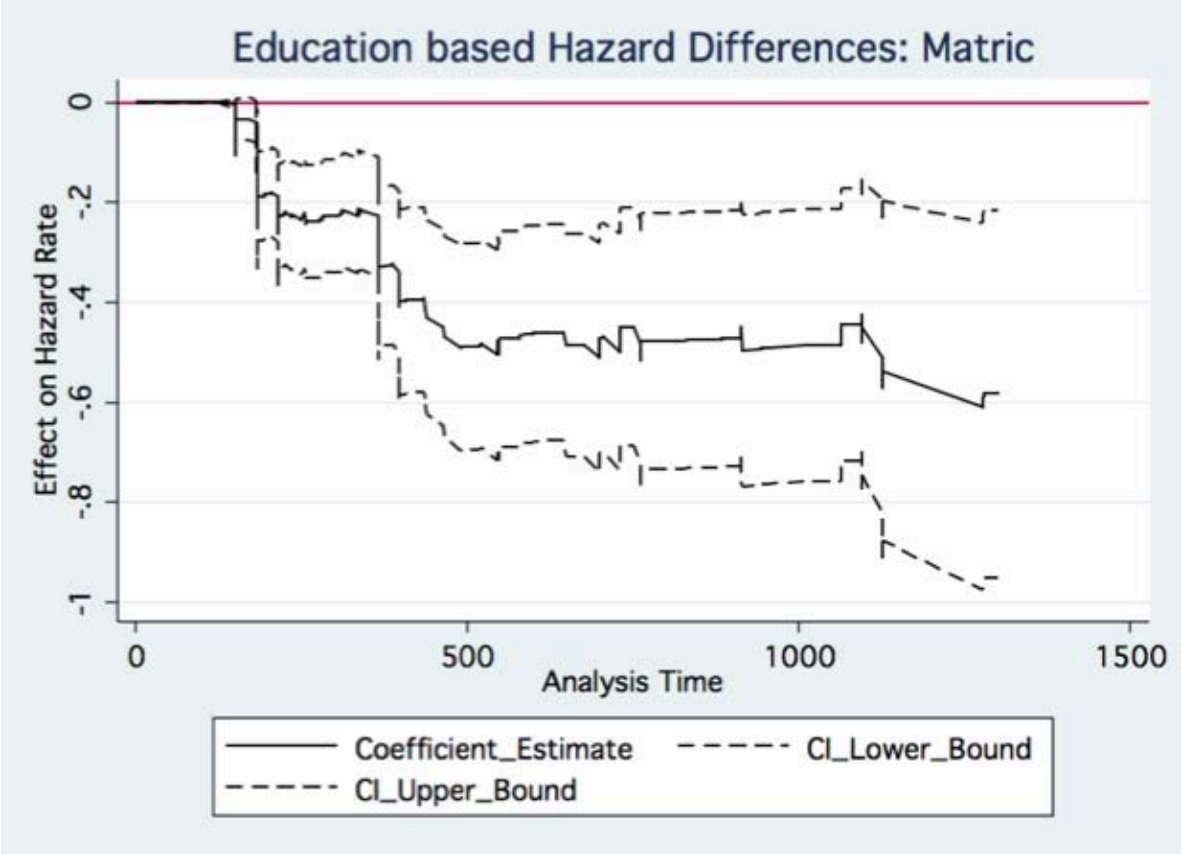


Figure 3.8.2.3b With Control Variables

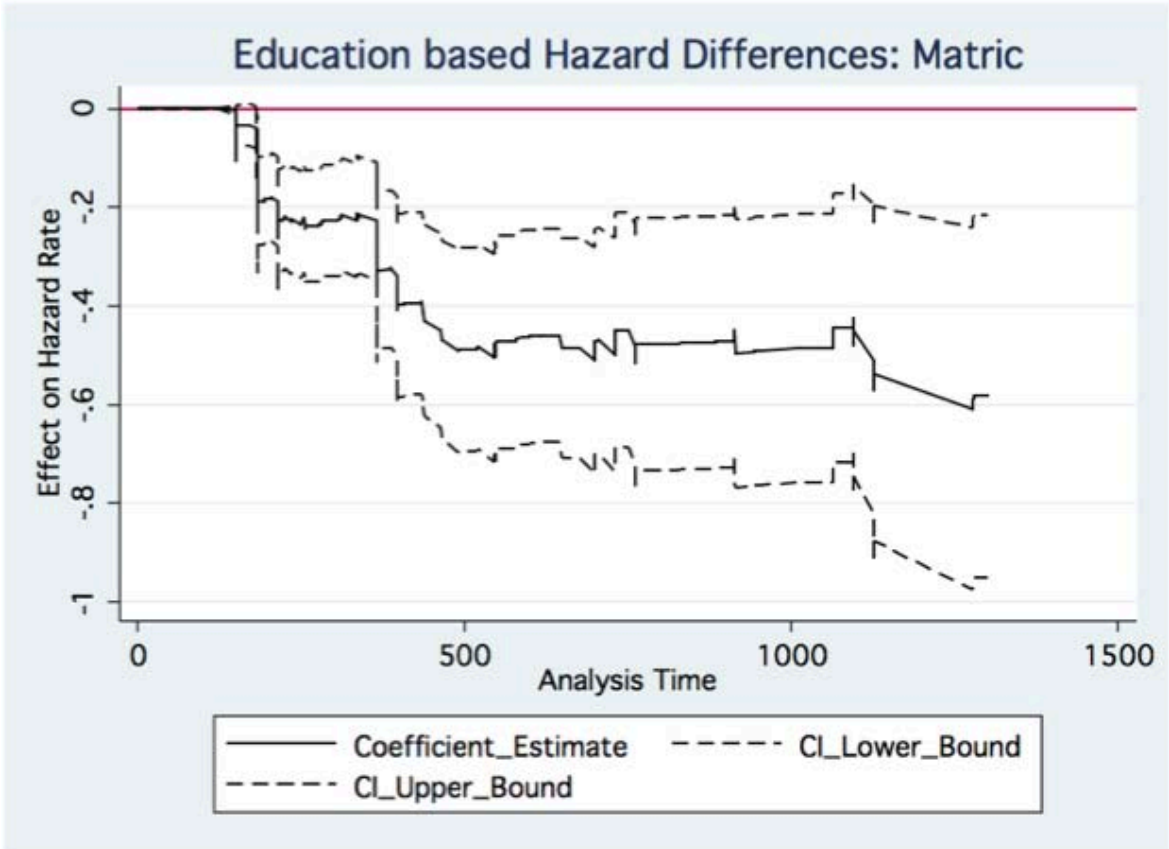


Figure 3.8.2.4.a No Control Variables

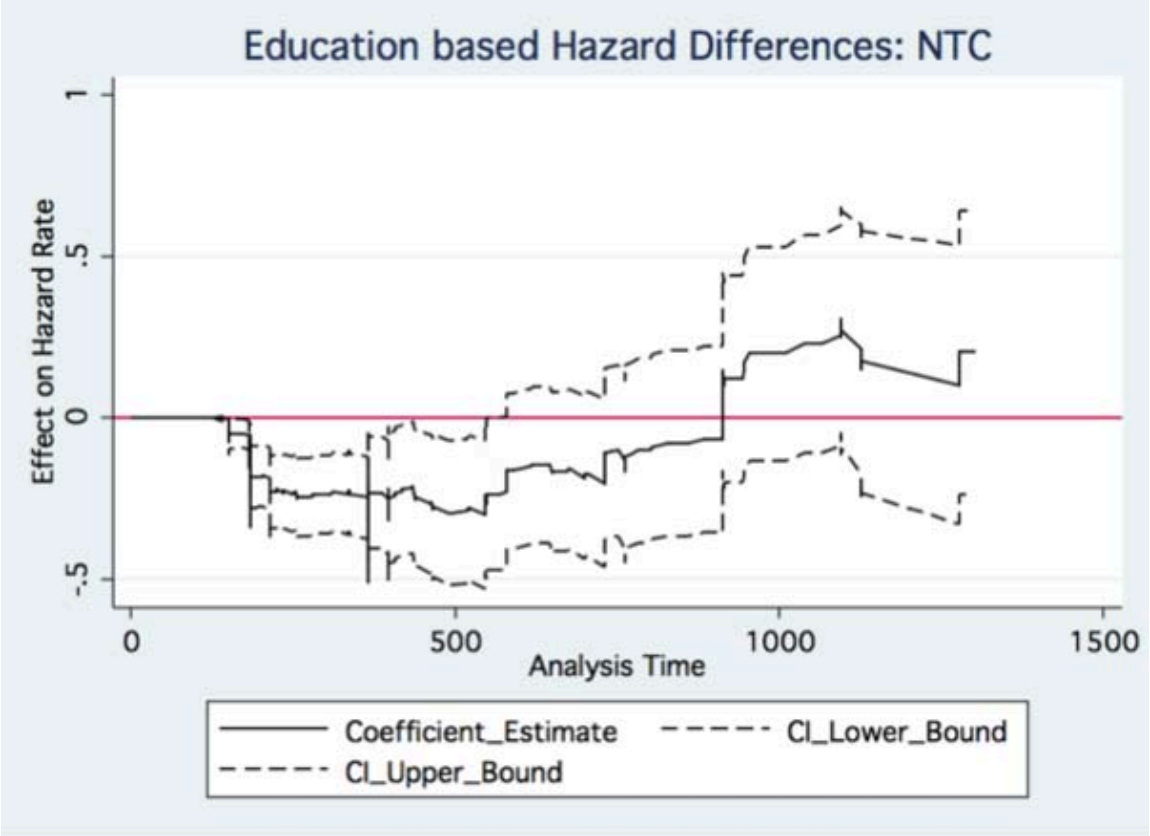


Figure 3.8.2.4b With Control Variables

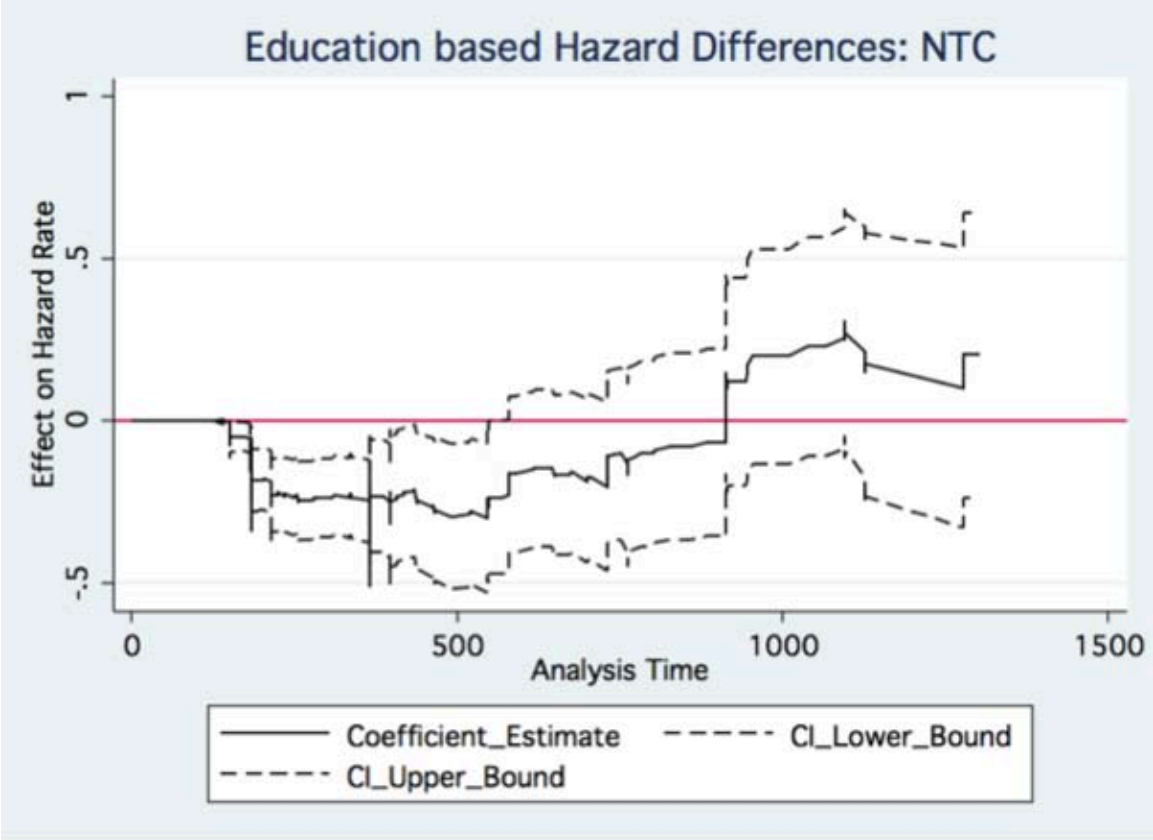


Figure 3.8.2.5.a No Control Variables

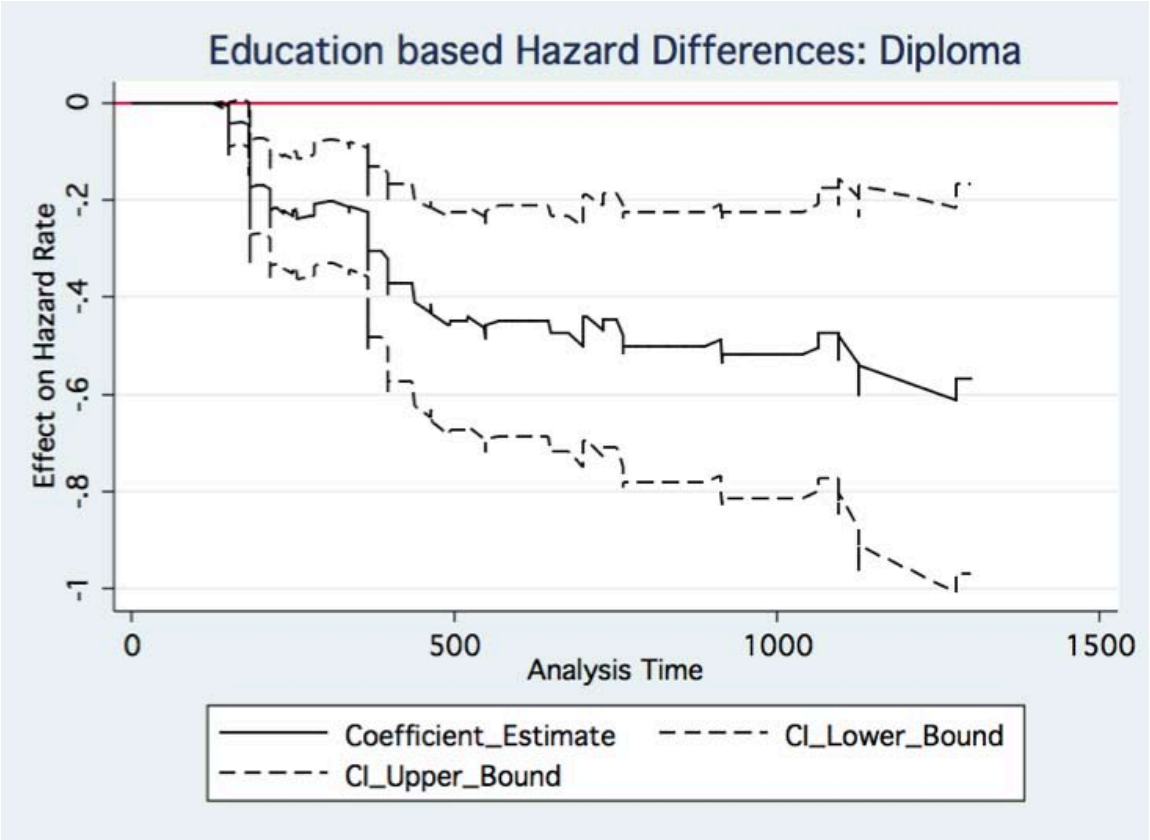
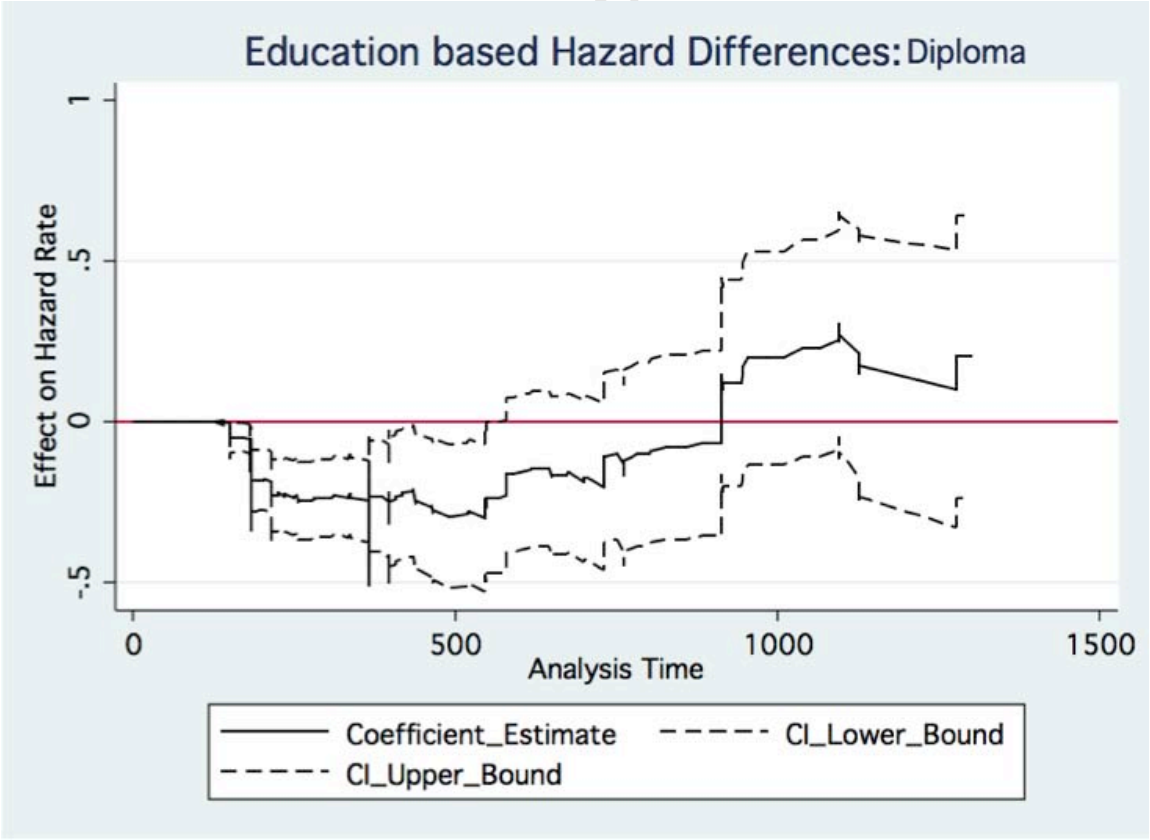


Figure 3.8.2.5b With Control Variables



The results for all of the education covariates indicate that, relative to University graduates, all other education levels apart from NTC graduates see an initial divergence in their hazard rates. Around 500 days of analysis time the hazard ratio stabilizes considerably. For NTC graduates the estimated coefficient is initially negative and statistically significant, but converges to zero over time. Controlling for other covariates yields largely similar results, however the difference between NTC graduates and University graduates is now statistically insignificant over the entire analysis period.

3.4.3 Rural-Urban Differences

Figure 3.8.3a No Control Variables

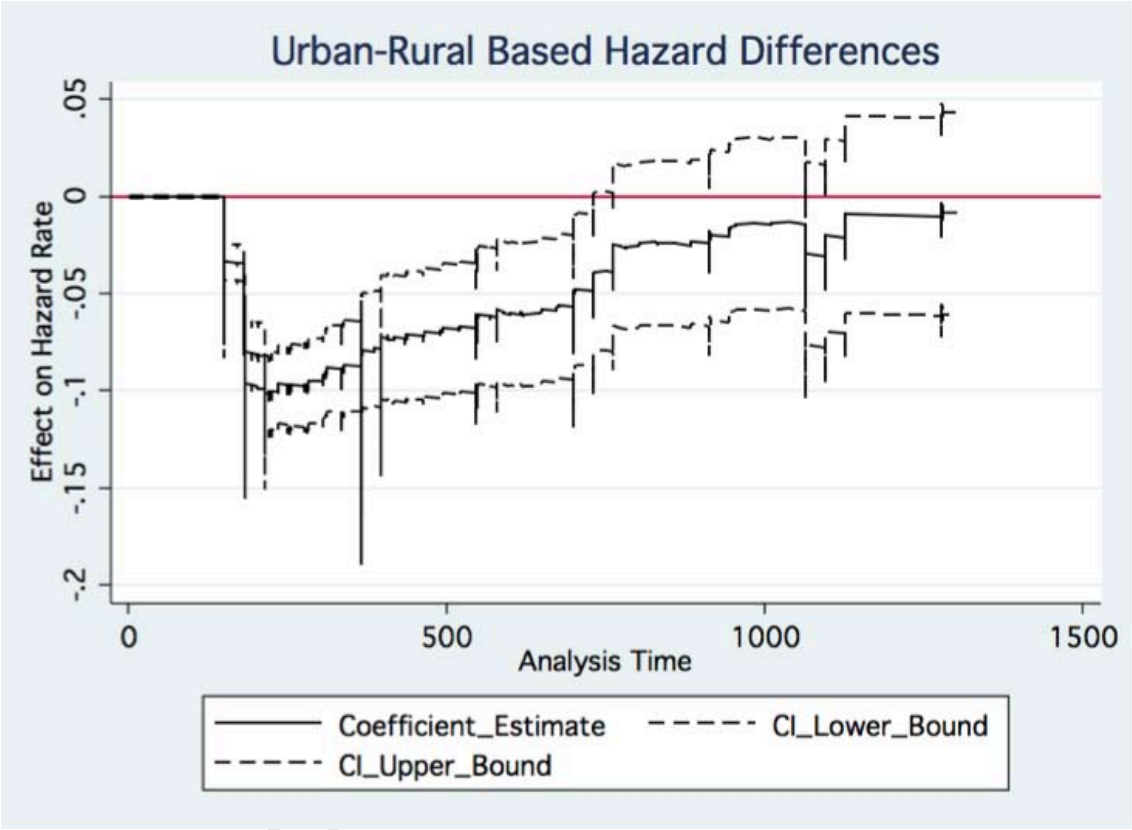
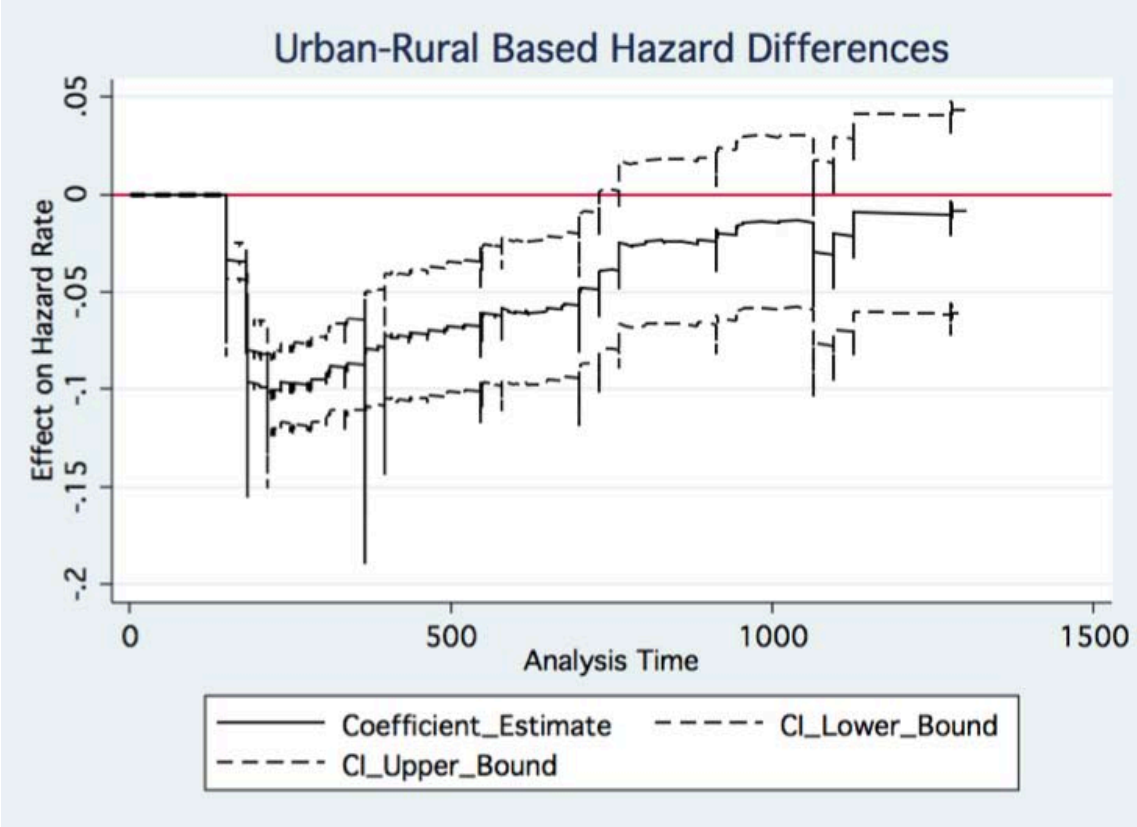


Figure 3.8.3b With Control Variables



When measuring for Rural-Urban differences previous sections suggest that while rural dwellers have a initial advantage in finding work the difference in hazard rates tends to zero over time. However, once the other covariates are controlled for the difference between urban and rural dwellers is more persistent and stable over time. This suggests that the effects of the urban residence on the job-search function is persistent and negative. An extended analysis of difference in urban hazard rates by province is provided in section 4.3.

3.4.4 Provincial Differences

Figure 3.8.4.1.a No Control Variables

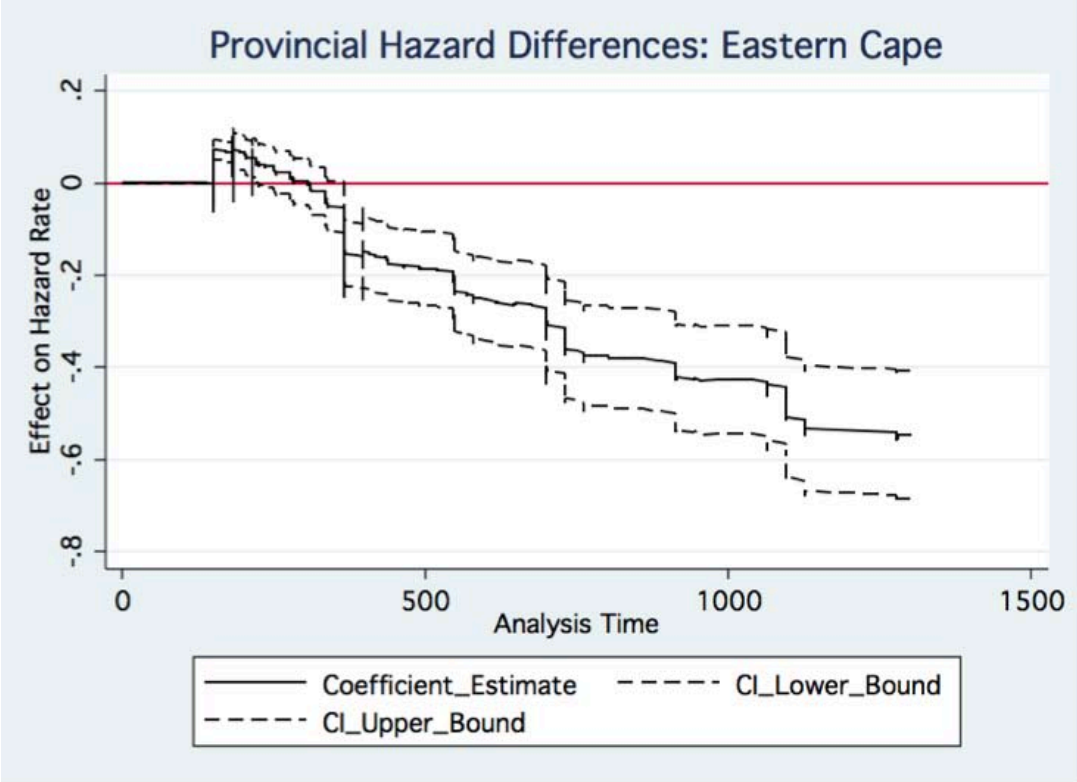


Figure 3.8.4.1b With Control Variables

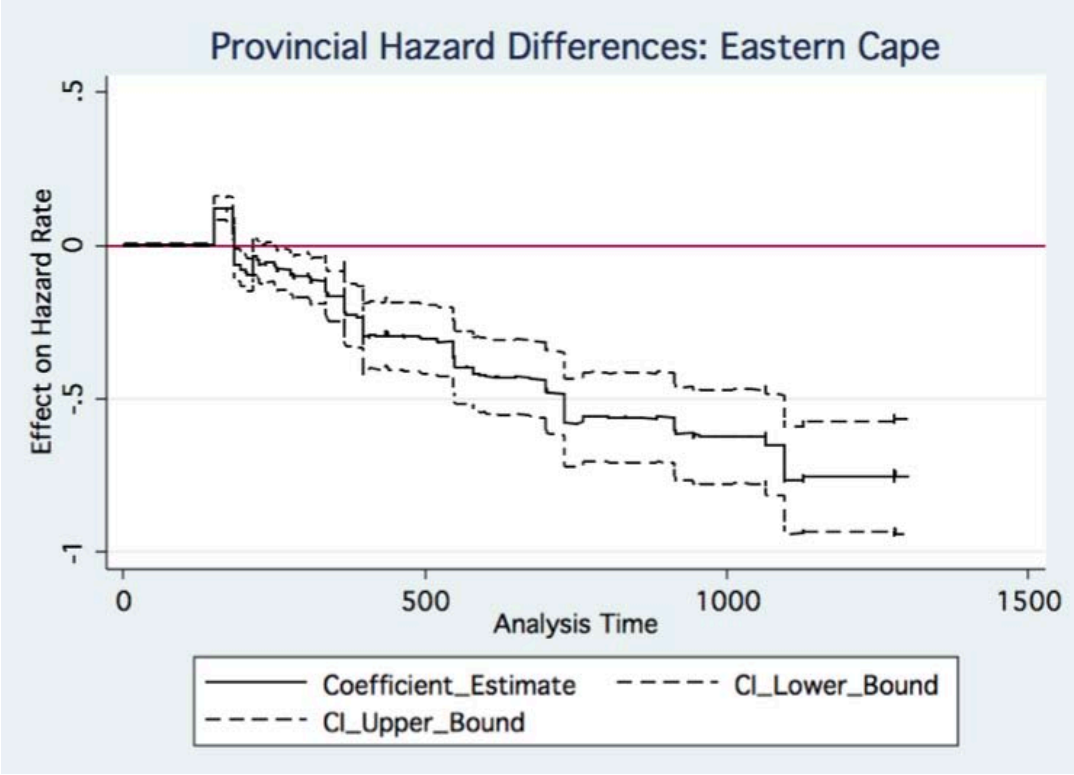


Figure 3.8.4.2.a No Control Variables

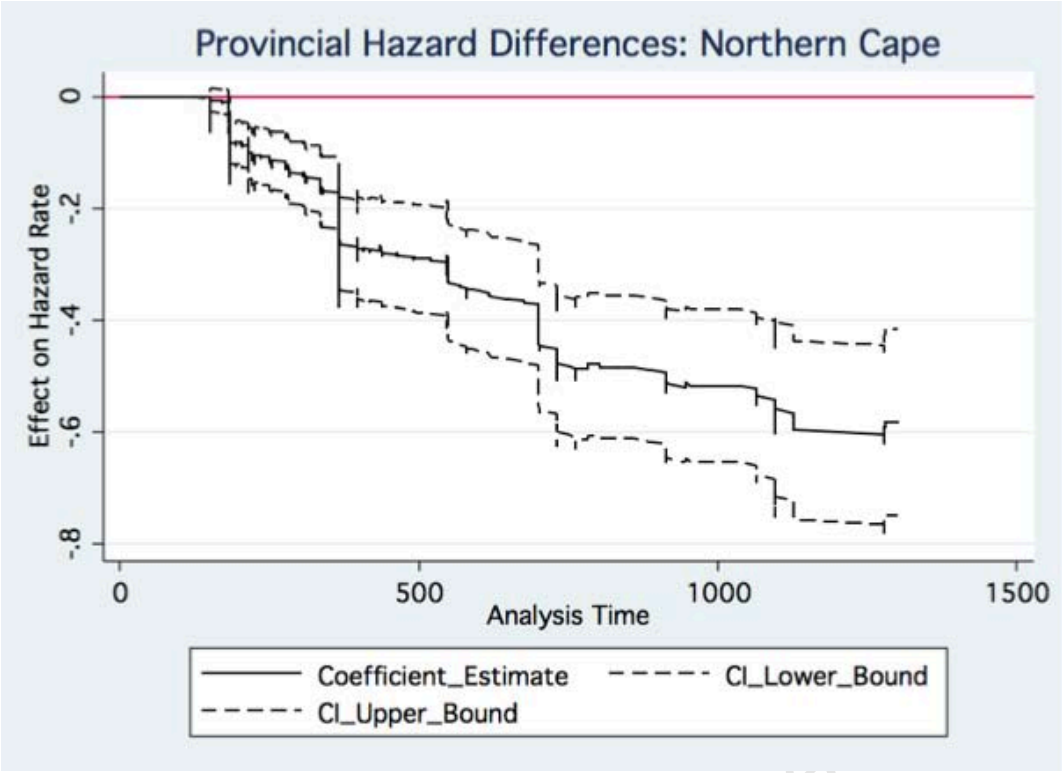


Figure 3.8.4.2b With Control Variables

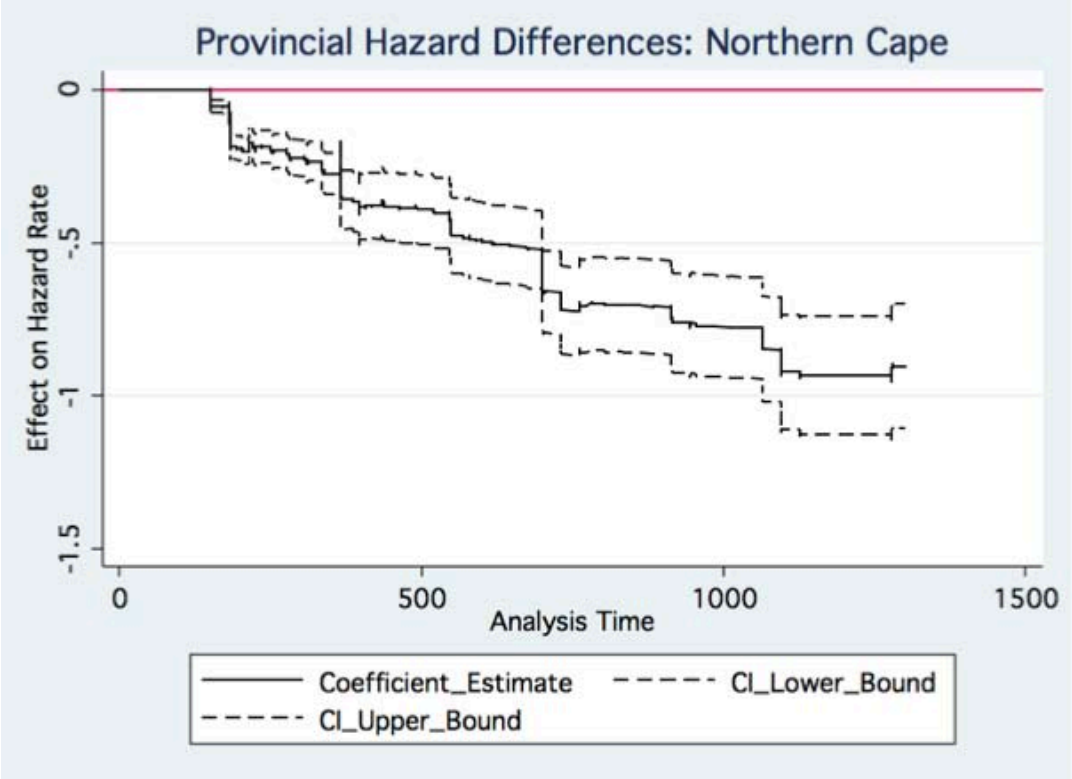


Figure 3.8.4.3.a No Control Variables

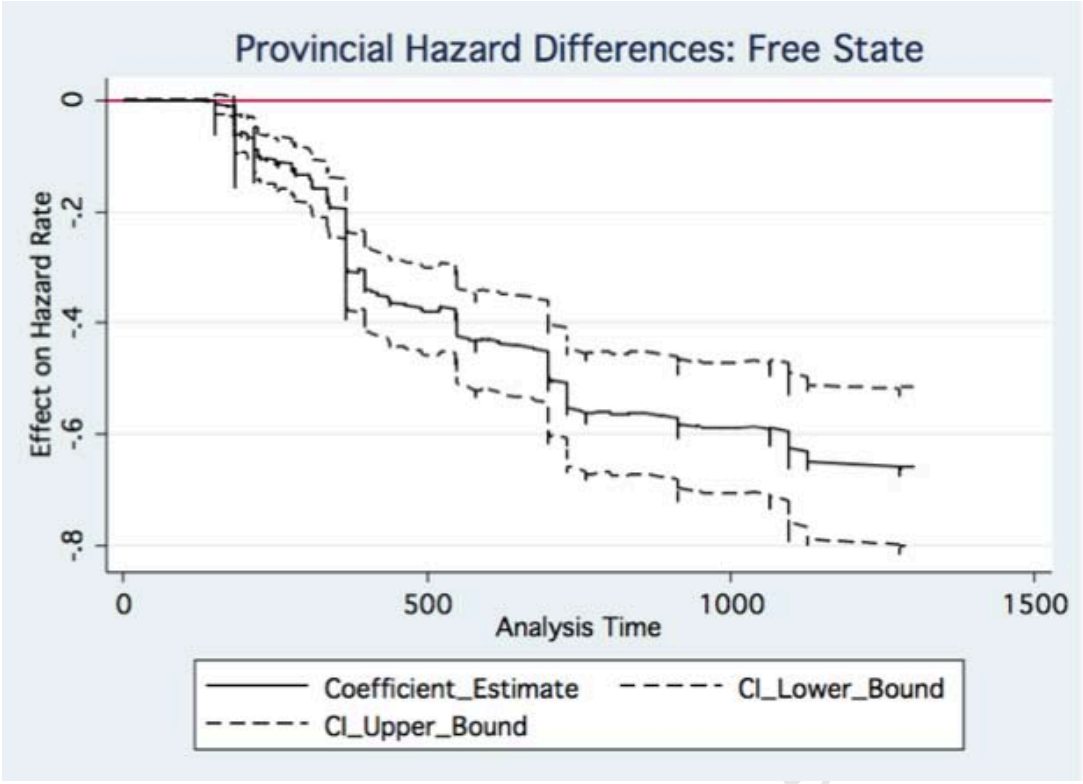


Figure 3.8.4.3b With Control Variables

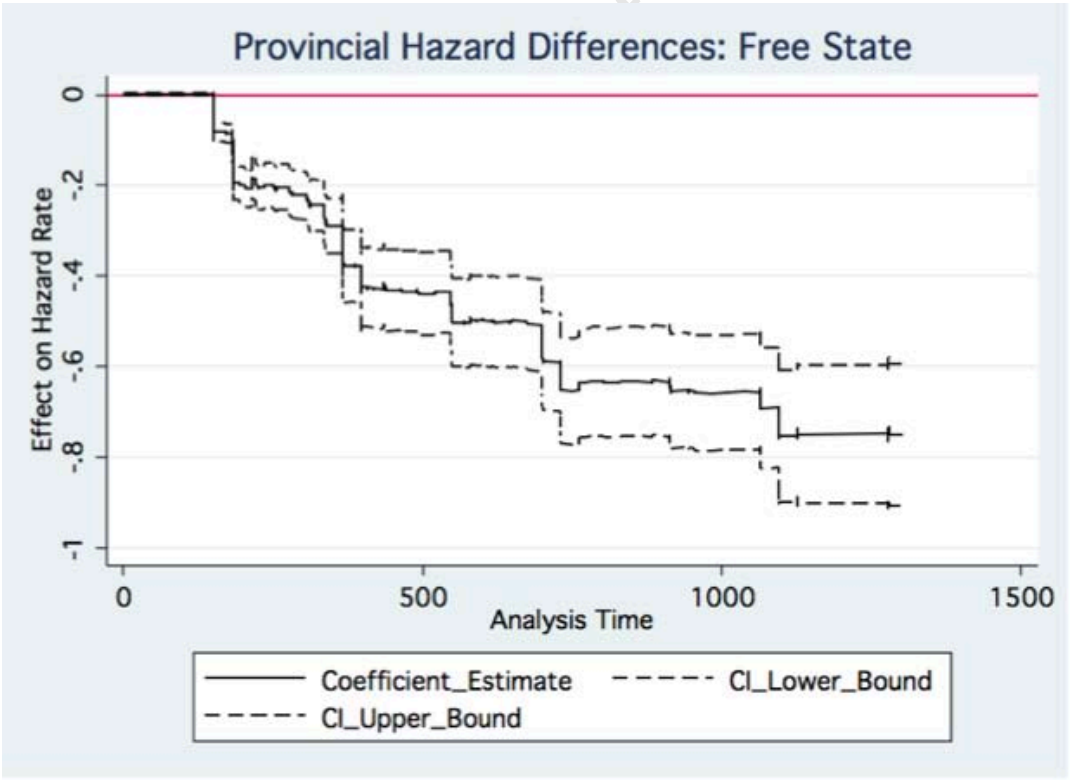


Figure 3.8.4.4.a No Control Variables

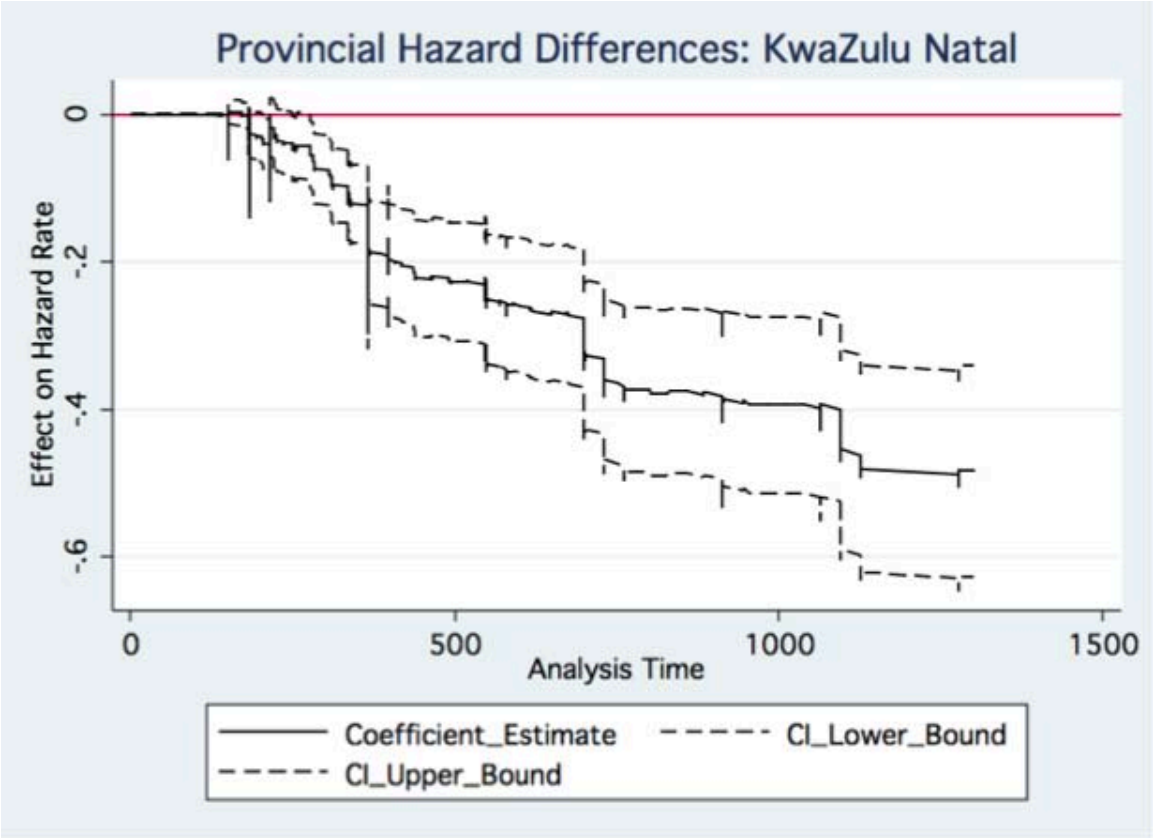


Figure 3.8.4.4b With Control Variables

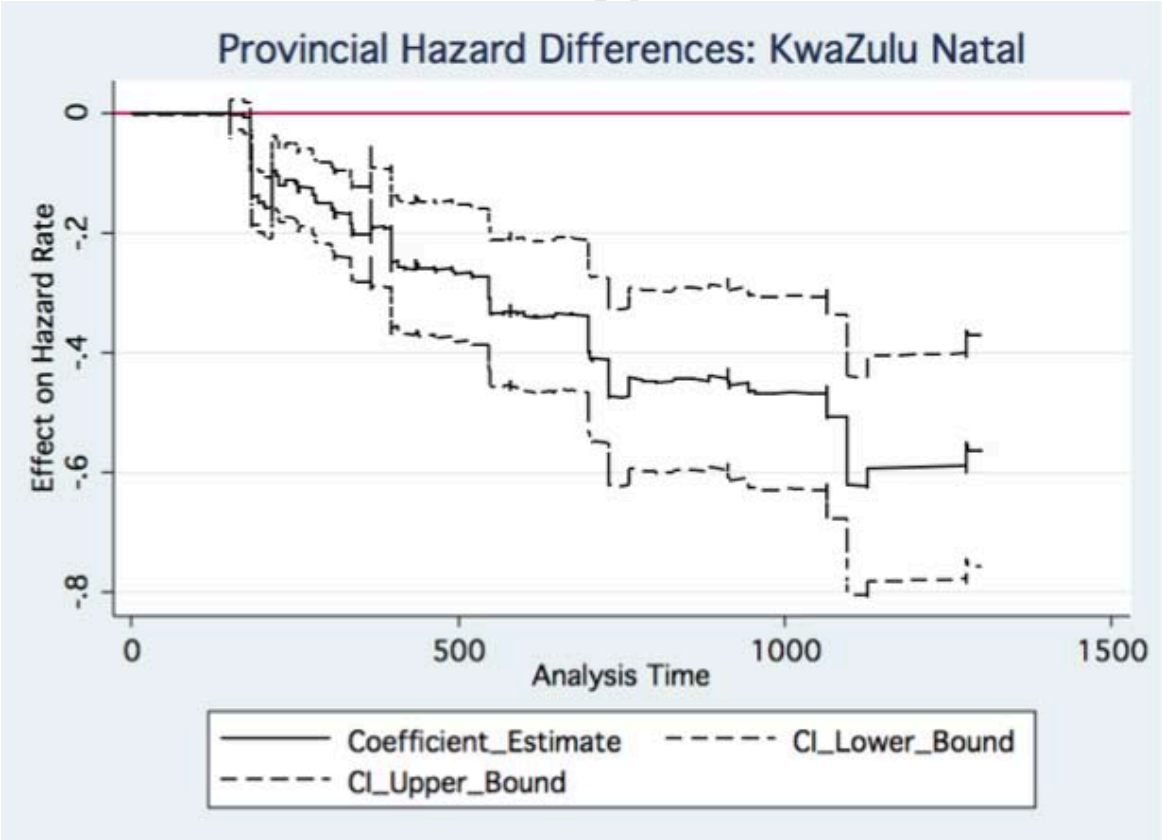


Figure 3.8.4.5.a No Control Variables

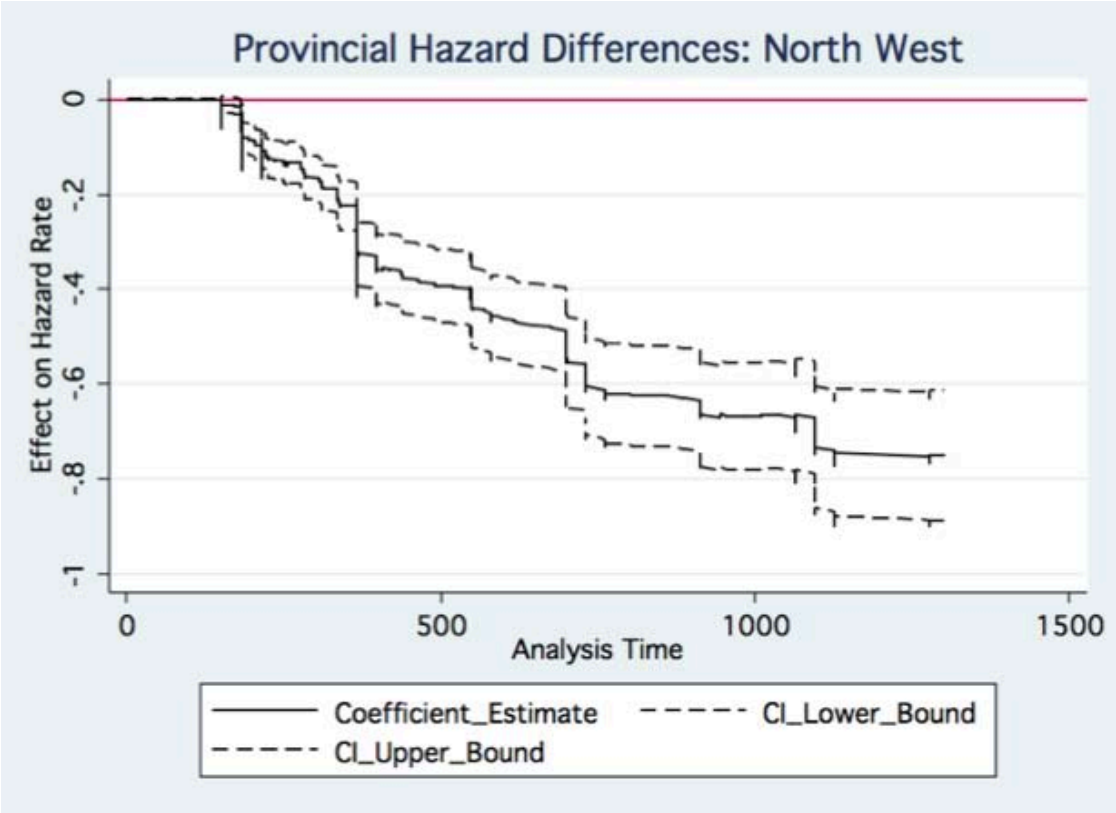


Figure 3.8.4.5b With Control Variables

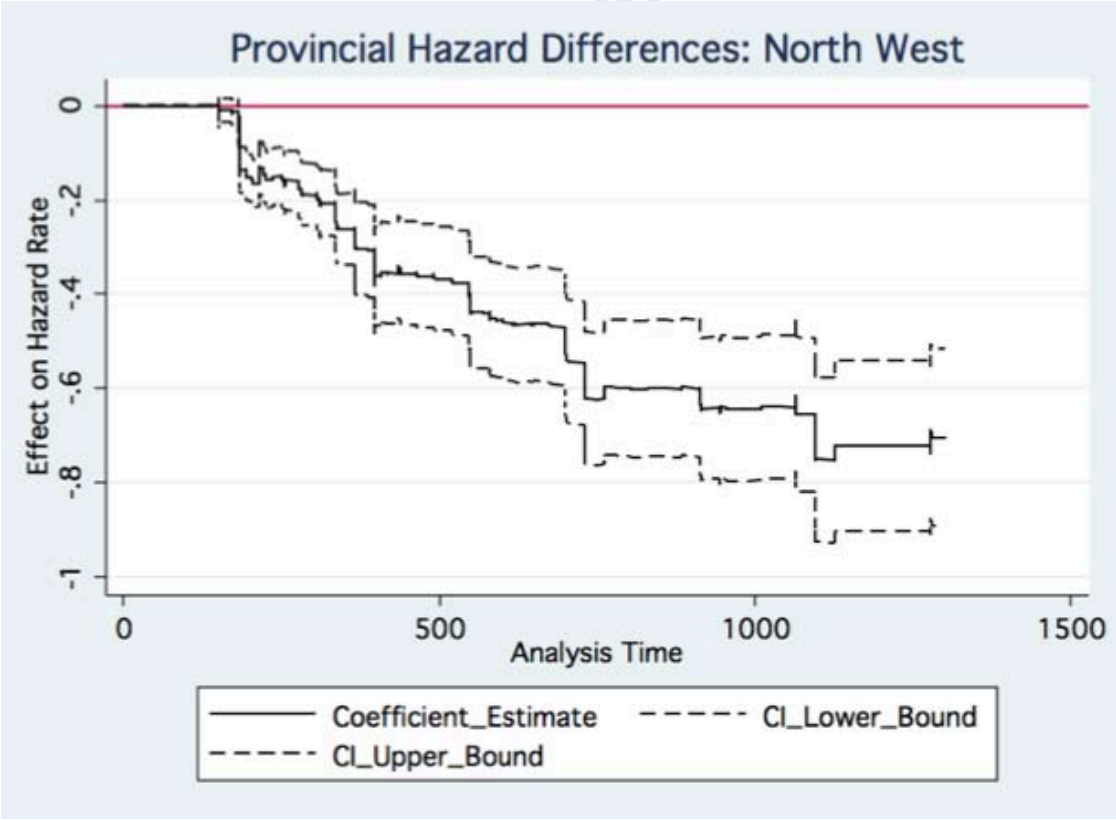


Figure 3.8.4.6.a No Control Variables

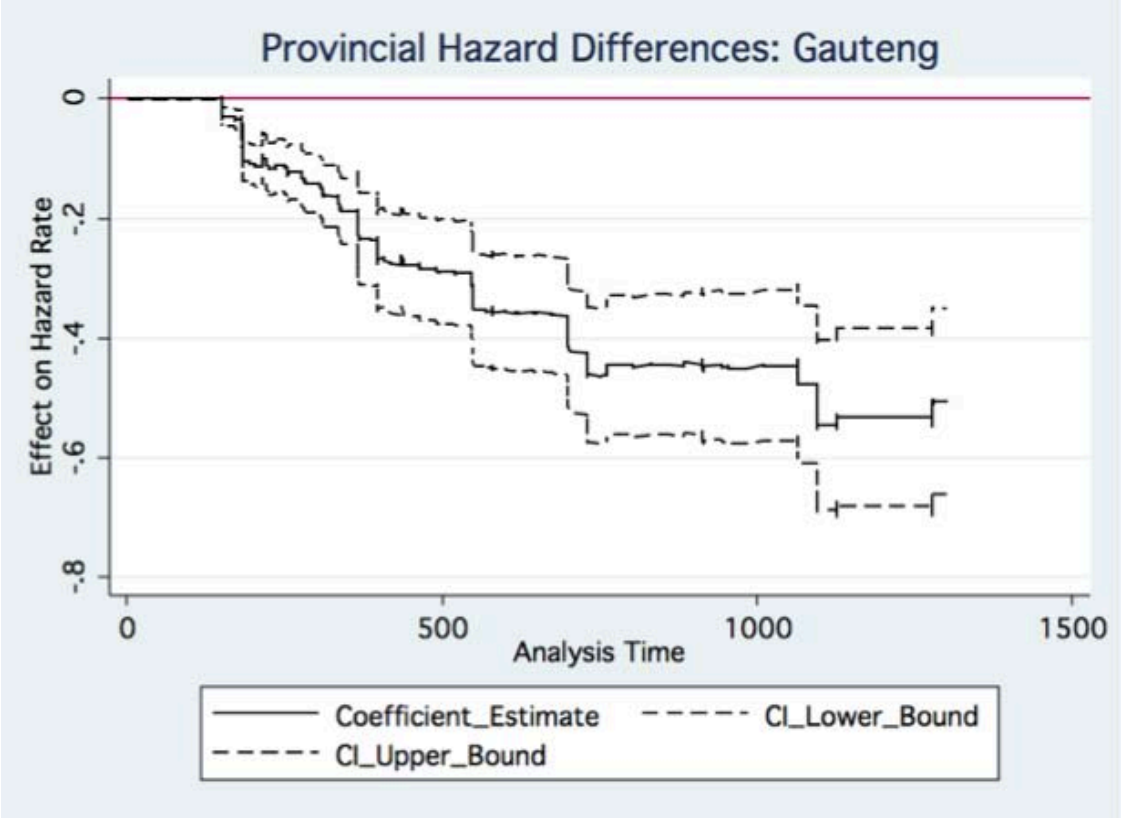


Figure 3.8.4.6b With Control Variables

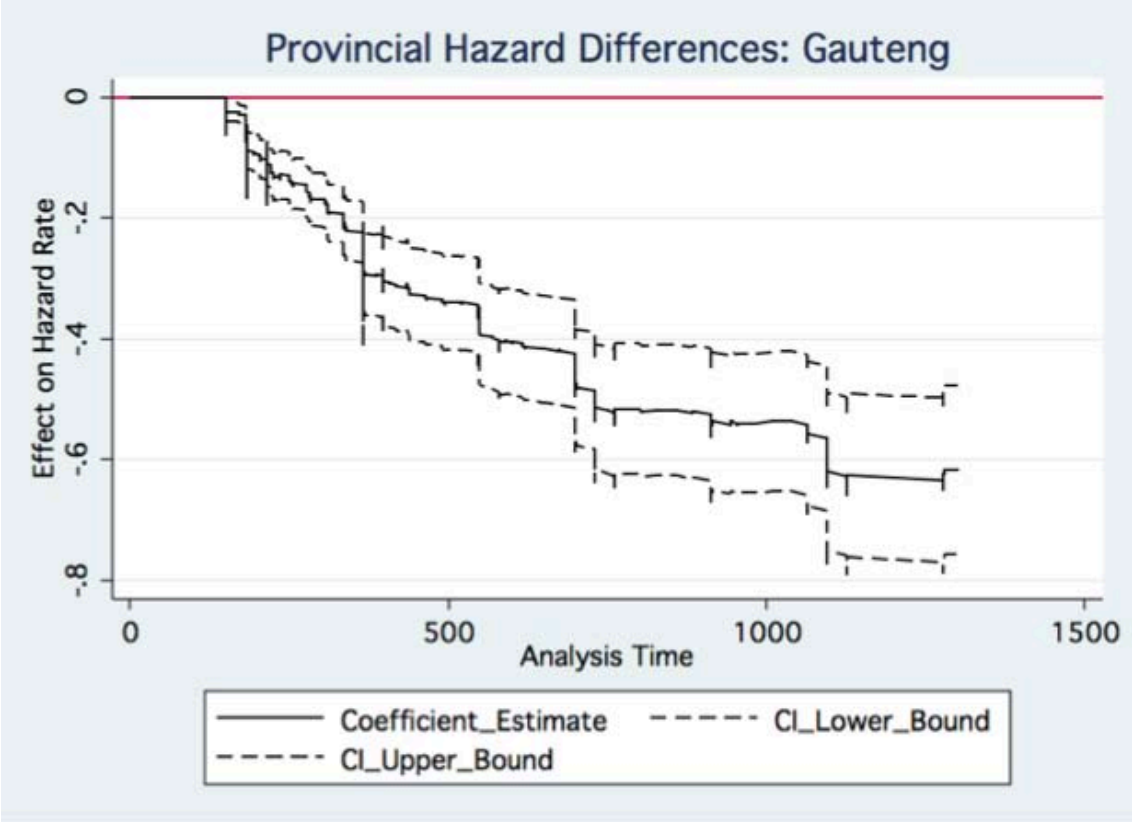


Figure 3.8.4.7.a No Control Variables

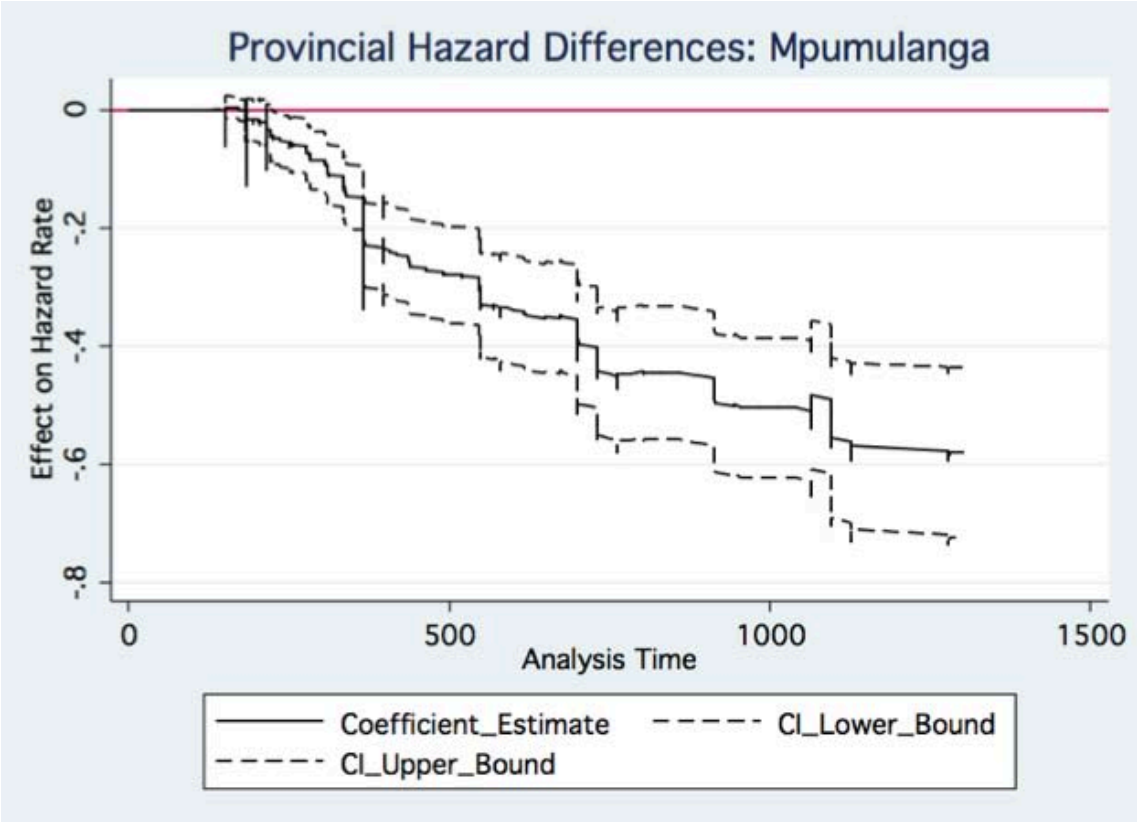


Figure 3.8.4.7b With Control Variables

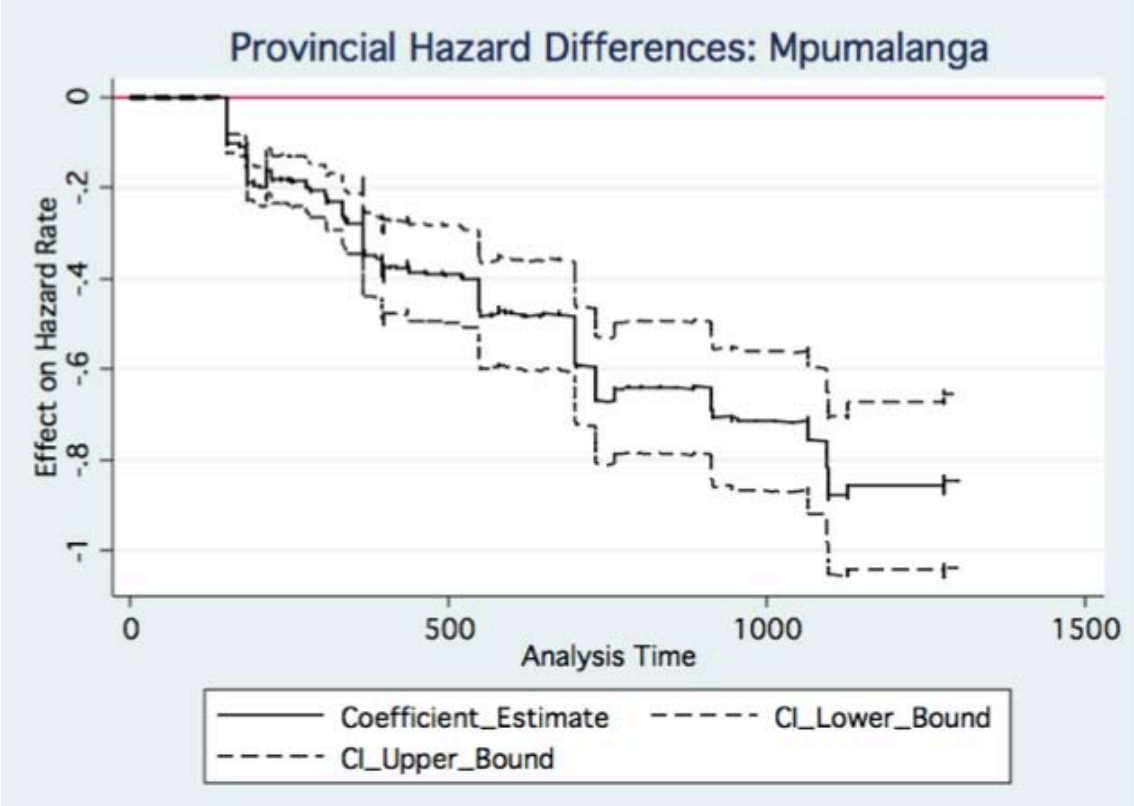


Figure 3.8.4.8.a No Control Variables

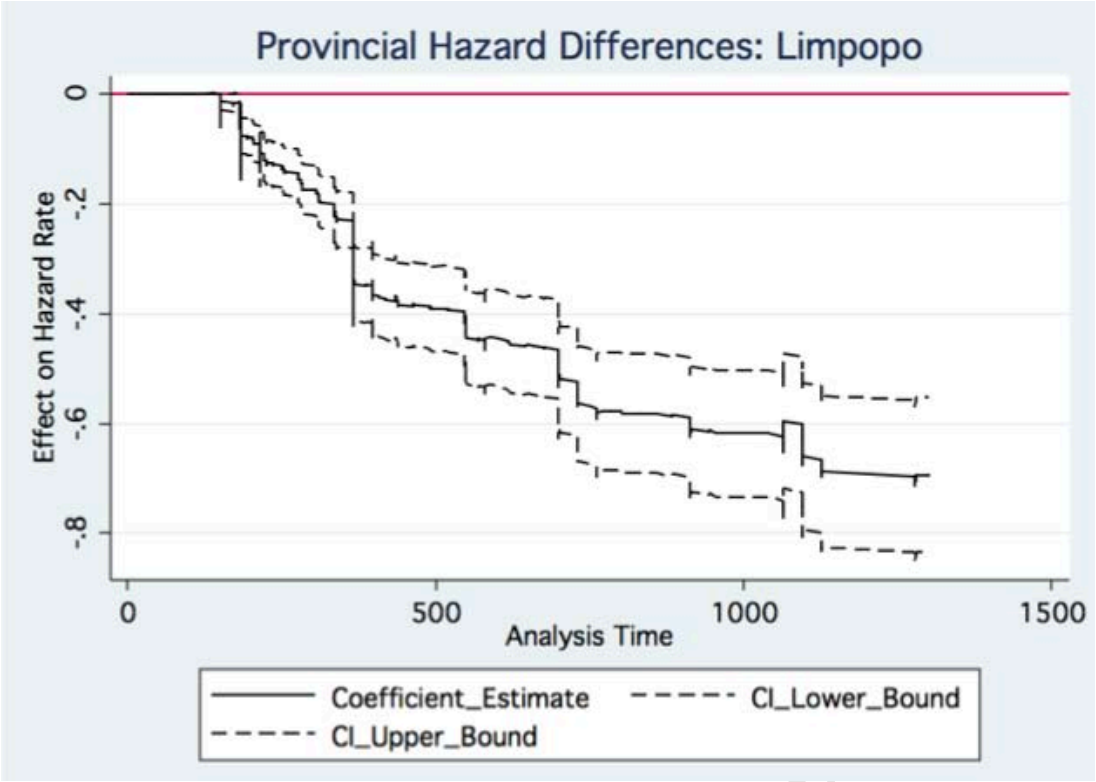
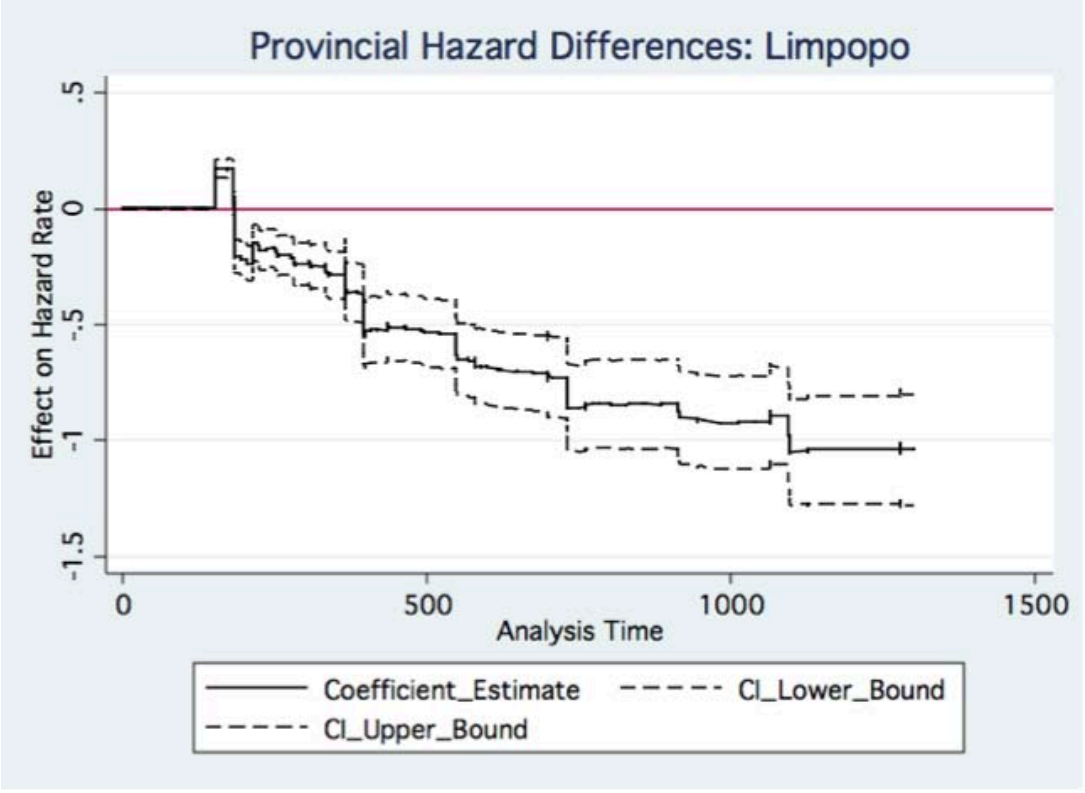


Figure 3.8.4.8b With Control Variables



The Aalen linear hazards model uniformly suggests divergence between the hazard rates of the Western Cape and all other provinces, in favour of the former. Note that this divergence persists over the entire period of analysis time considered.

3.4.5 Racial Differences

Figure 3.8.5.1.a No Control Variables

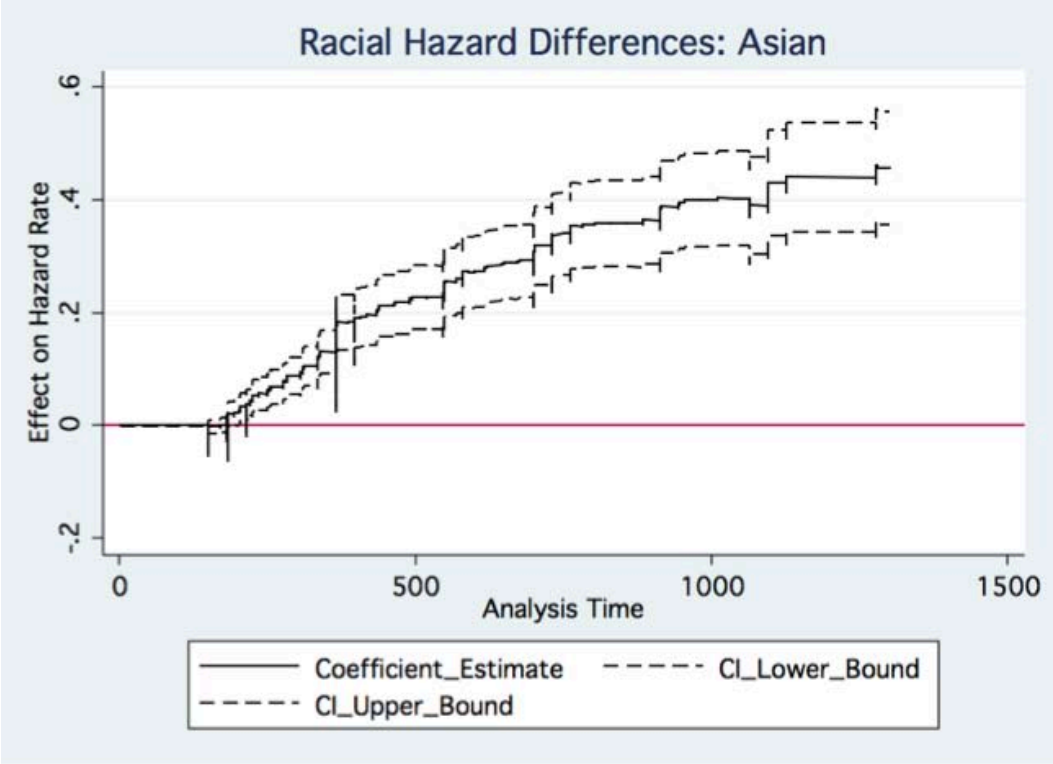


Figure 3.8.5.1b With Control Variables

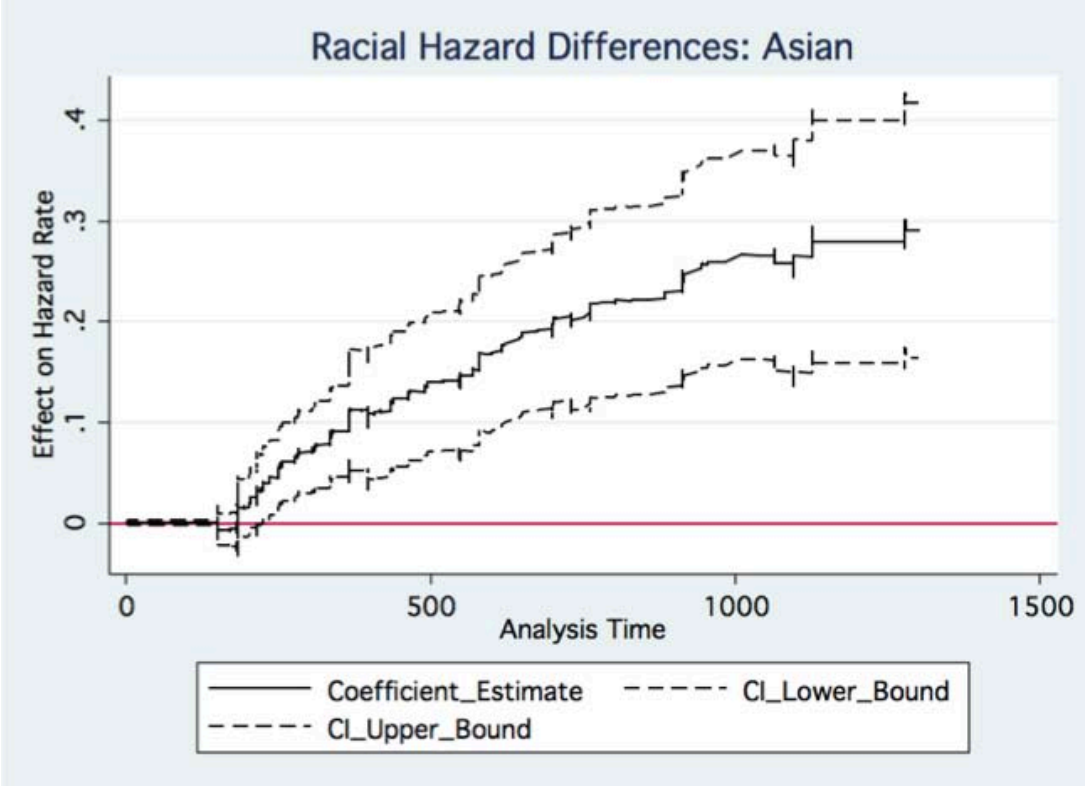


Figure 3.8.5.2.a No Control Variables

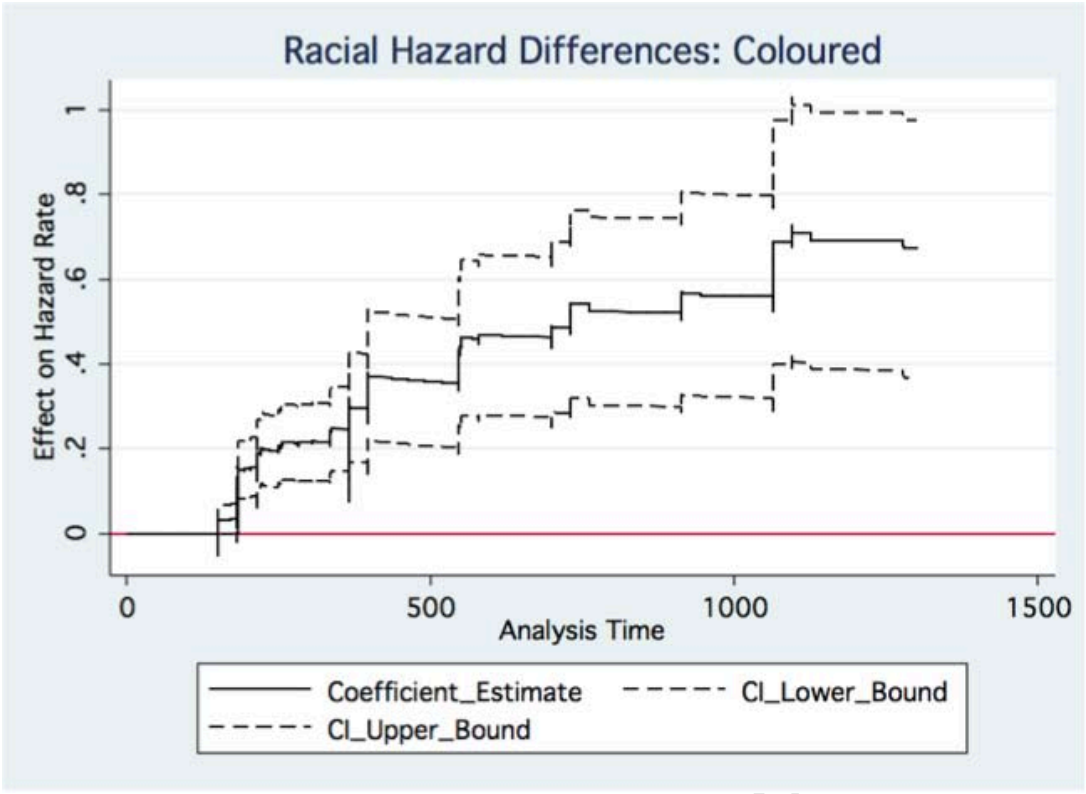


Figure 3.8.5.2b With Control Variables

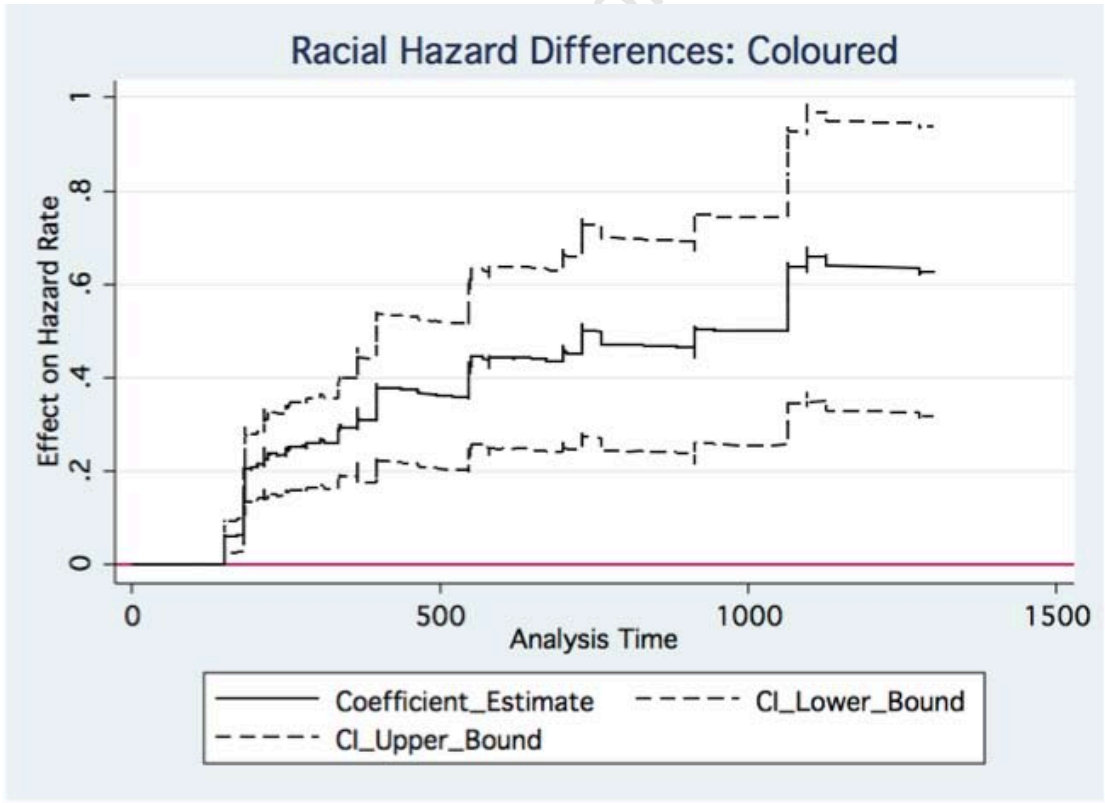


Figure 3.8.5.3.a No Control Variables

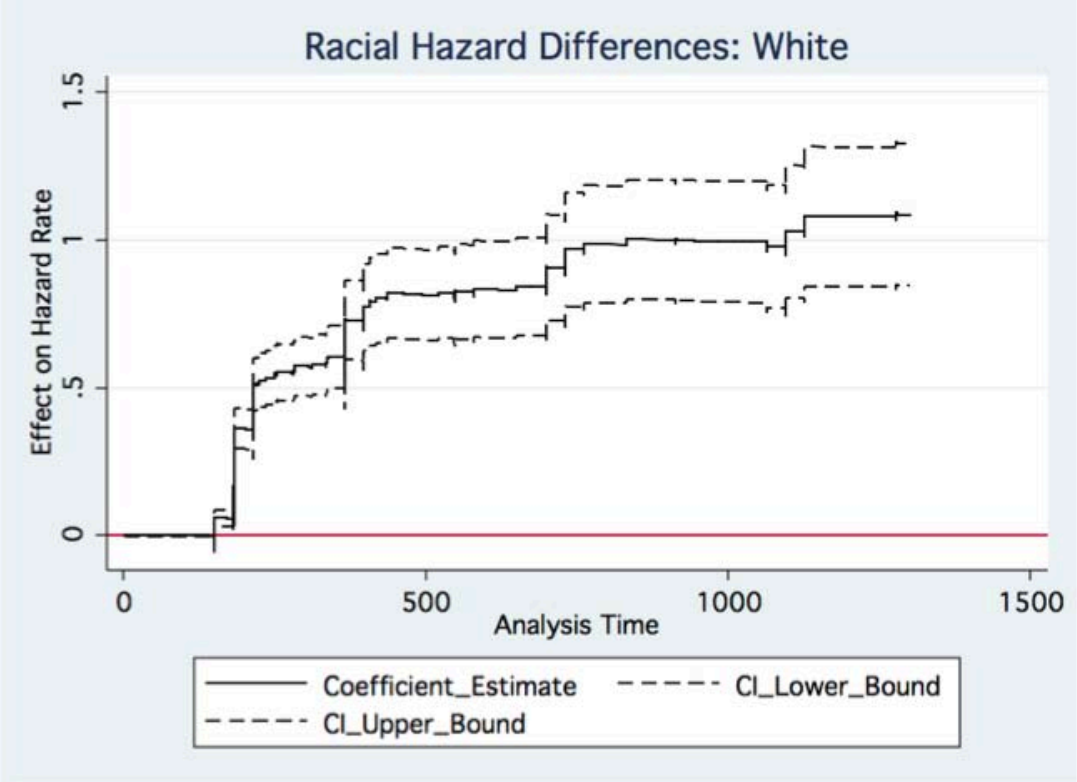
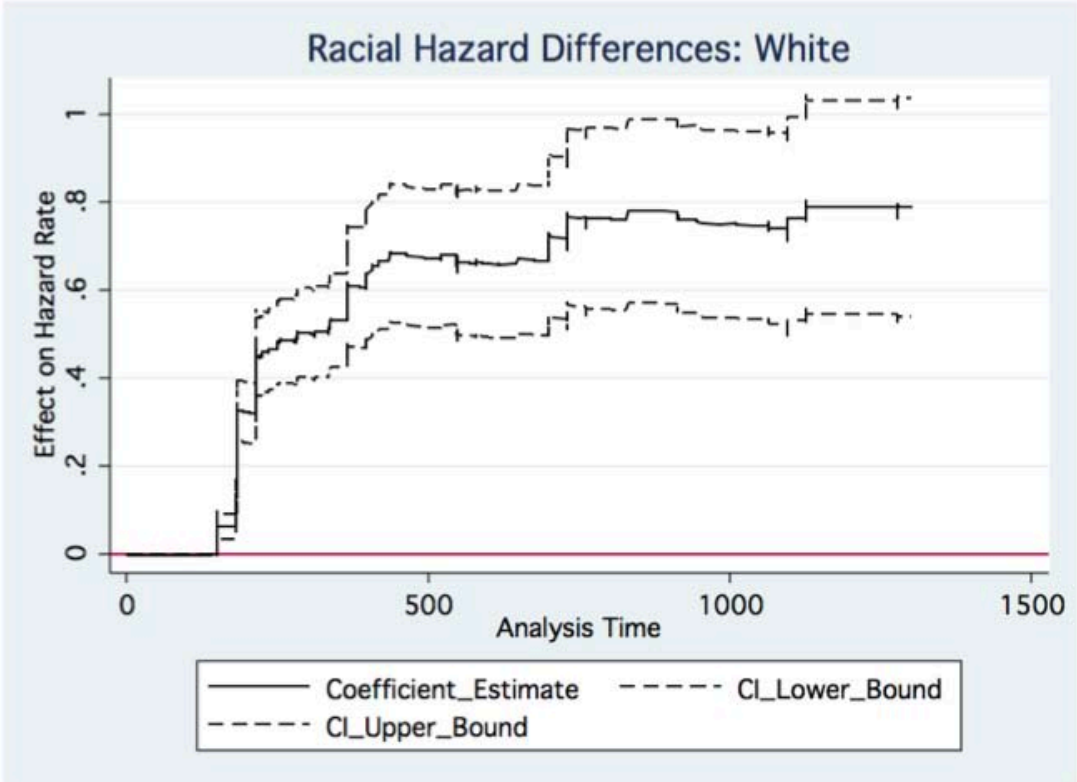


Figure 3.8.5.3b With Control Variables



As with the provincial indicators, the Aalen linear hazards model uniformly suggests divergence between the hazard rates of the different races. Once I control for other covariates the racial hazard rates do seem to stabilize somewhat for the white indicator variable after 500 days of analysis time.

4 Extending the Analysis

So far the analysis has yielded several interesting results, to improve my understanding of these results I will extend my analysis in three directions in an attempt to answer 3 questions:

1. To what extent is my analysis of the long-term unemployment driven by discouraged workseekers in the sample?

2. To what extent are the differences in labour market experiences between educational groups driven by differences in schools quality amongst the different racial groups?

3. How do the urban and rural sections of the nine provinces compare to each other?

4.1 Discouraged Work-Seekers in the Sample

Table 4.1 Observations in the Sample Associated with Failure

	No-Failure	Failure
Non-Discouraged	11,406	4,206
Discouraged	6,269	0

None of the discouraged work seekers transitioned out of unemployment. Sadly the job market experience of discouraged work-seekers seem to confirm Seeking´s (2003) classification of the chronic unemployed as an economic underclass, excluded from employment opportunities. It may therefore be worthwhile to repeat the entire analysis excluding discouraged work-seekers and focusing on those officially defined as unemployment.

4.2 Race and Education in the South African Labour Force

One surprising results from my initial analysis was that a completed secondary education was not associated with any labour market benefit. However, this result obscures the large historical inequalities in the South African education system. Particularly, the educational system under apartheid was aimed primarily at one level of education to the country´s white minority and a vastly inferior level of education to the non-white majority. In non-white schools these policies where characterized by high pupil-teacher ratios, poorly qualified teachers and low levels of funding (Fedderke *et al.* 1998).

A natural extension of my analysis is to compare labour market experiences for different educational levels by race. To complete this analysis I interact the education covariate by a racial indicator variable, and estimate a Cox-PH model. Note that this regression does not control for other

covariates and is therefore only a comparison of labour market experiences by demographic subgroups. As a baseline I use white matriculants who represent the so-called model-C schools.

Table 4.2 Race and Education in the South African Labour Market: A Cox PH Analysis

	Haz. Ratio	Std. Err.	P>z	[95% Confidence	Interval]
No-high school	3.244	1.475	0.010	1.331	7.909
Some high school	0.790	0.108	0.086	0.604	1.034
NTC	1.515	0.278	0.023	1.058	2.170
Diploma	0.831	0.157	0.329	0.573	1.205
University degree	1.278	0.215	0.144	0.920	1.777
African	0.266	0.025	0.000	0.221	0.320
No-high school	0.458	0.210	0.088	0.186	1.123
Some high school	1.380	0.206	0.030	1.031	1.849
NTC	1.257	0.257	0.263	0.842	1.876
Diploma	1.253	0.290	0.329	0.796	1.971
University degree	1.284	0.297	0.279	0.817	2.019
Asian	0.578	0.070	0.000	0.456	0.733
No-high school	0.298	0.139	0.010	0.119	0.746
Some high school	0.982	0.171	0.916	0.698	1.381
NTC	0.868	0.219	0.574	0.530	1.422
Diploma	1.073	0.394	0.848	0.522	2.205
University degree	1.245	0.500	0.586	0.567	2.736
Coloured	0.705	0.108	0.022	0.522	0.951
No-high school	0.516	0.280	0.223	0.178	1.495
Some high school	0.801	0.205	0.386	0.486	1.322
NTC	0.838	0.287	0.606	0.429	1.639
Diploma	1.631	0.660	0.227	0.737	3.606
University degree	1.383	0.637	0.481	0.561	3.409

Note that comparing hazard rates using the interaction term is slightly complicated. The HR between a white matric and a matric of any other race is simply equal to the estimated hazard ratio of the racial indicator. The hazard ratio between black and white matrics is therefore 0.266. However, if I wish to compare the white matrics to another educational group of another race I need to multiply the hazard ratios for all the applicable variables. For example the hazard ratio between white matrics and asian university graduates will be given as

HR(Asian university: white matric)

= HR(university:matric)*HR(asian:white)*HR(Asian*university:White*matric) = 1.278*0.578*1.245

= 0.91966.

This result follows from the functional form of the Cox PH model.

Similarly I could calculate the hazard ratio between white matriculants and coloured NTC graduates as 0.895. To calculate the hazard ratio of Asian university graduates and coloured NTC graduates I calculate it as 0.91966/0.895=1.02755. Statistical significance for a given race indicator variables suggest that matrics of said race face a different hazard rate than white matrics. Statistical significance for the coefficients of the interaction terms in the estimated Cox model suggests that a subgroup differ statistically, both from Whites with the same educational level and matrics of the same race.

The results of table 4.2 clearly indicate large statistically significant differences in hazard rates for matrics of different races. Of these african matrics fare the worst by far, with a hazard rate equal to only 26.6% that of white matrics. For levels of education greater than matric most subgroups do not perform in a manner out of line with their racial grouping and level of education, as is borne out by the lack of significance for most of the interaction terms. Once accounting for racial differences a matric is found to offer a similar amount of labour market benefit than a diploma or university degree, but less than an NTC qualification.

Table 4.3 Sectoral Employment by Educational Attainment

	Formal Economy	**Informal Economy**
No-High school	8,052	4,806
Some High School	11,268	3,881
Matric	9,961	1,477
NTC	3,227	376
Diploma	4,581	266
Degree	4,344	174

Surprisingly those with an incomplete high school education consistently outperform matrics for all

racial groupings. One potential explanation for this could be the lower reservation wages amongst those without completed matric education. In the sample those with less than a matric education are much more likely to be observed working in the informal sector as opposed to the formal sector (see table). The informal sector is characterized by lower wages and an absence of restrictive labour legislation. The discrepancy in hazard rates between matric and non-matric may therefore be an indication that the informal sector is more labour absorbing than the formal sector.

4.3 Rural and Urban Labour Market Experiences in South Africa´s 9 Provinces

Table 4.4 Rural and Urban Labour Market Experiences in South Africa´s 9 Provinces: A Cox PH Analysis

	Haz. Ratio	Std. Err.	P>z	[95% Conf.	Interval]
Province					
Eastern Cape	0.677	0.064	0.000	0.562	0.814
Northern Cape	0.423	0.068	0.000	0.309	0.580
Free State	0.307	0.040	0.000	0.238	0.397
KwaZulu Natal	0.401	0.040	0.000	0.330	0.486
North West	0.269	0.029	0.000	0.217	0.334
Gauteng	0.328	0.119	0.002	0.161	0.667
Mpumalanga	0.407	0.042	0.000	0.333	0.497
Limpopo	0.284	0.028	0.000	0.233	0.346
Urban	0.521	0.048	0.000	0.435	0.624
Eastern Cape	0.817	0.100	0.098	0.644	1.038
Northern Cape	1.463	0.271	0.040	1.018	2.104
Free State	1.959	0.296	0.000	1.457	2.634
KwaZulu Natal	2.012	0.245	0.000	1.585	2.555
North West	1.981	0.275	0.000	1.509	2.600
Gauteng	1.839	0.677	0.098	0.894	3.783
Mpumalanga	1.565	0.207	0.001	1.207	2.028
Limpopo	2.043	0.314	0.000	1.512	2.761

Interpretation of the urban interaction terms in the Cox PH model in table 4.3 follow in exactly the same fashion as the education interaction terms of section 4.2. Note that Limpopo, the Free State, KwaZulu Natal and the North West province all show marginally higher hazard rates in urban as opposed to rural settings. On the other end of the spectrum, the Eastern-, Western- and Northern Cape

all showed significantly lower hazard rates for urban dwellers. All the interactions terms and provincial indicator variables had statistically significant coefficient estimates indicating large-scale variation in labour-market experiences in South Africa based on geographical location. A complete analysis economic differences of the rural-urban sectors of each of South Africa´s nine provinces would clearly be instructive as to the sources of these differences. However, such an analysis lies outside the scope of this thesis.

5. Conclusion: Who has the best/worst labour market experience?

This thesis has attempted to estimate the differences in labour market experiences amongst unemployed South Africans. I set out to evaluate who had the best and worst labour market experiences in SA, the results of the survival analysis has suggested a nuanced answer wherein geographical location, race and educational levels all played an important role.

Having established the complexity of the underlying job search market in South Africa I will proceed, cautiously, to ranking labour market experience. My ranking has been done by covariates in isolation, which is intended primarily as a measurement of differences between subgroups.

Table 5.1 Ranking Labour Market Experiences

Gender	Education	Rural-Urban	Province	Race	Age
Female	University	Rural	Western Cape	White	65
Male	NTC	Urban	Eastern Cape	Asian	51
	No-high school		KwaZulu Natal	Coloured	
	Diploma		Mpumalanga	African	
	Matric		Northern Cape		
	Some high school		Gauteng		
			Free State		
			Limpopo		
			North West		

Table 5.1 uses the Cox PH model to rank labour market experience from best to worst. The Cox-PH model suggests that a 51-year-old African male living in the North West Province with an incomplete high school education experiences the worst labour market conditions. Conversely, a white female university graduate living in the rural Western Cape should experience the best labour market conditions, curiously my ideal labour market candidate is also a retiree.

If these hypothetical labourers seem unlikely, it is because their composition reflect broad trends more than a composite of individual characteristics. They do however, serve the purpose of highlighting that it is amongst the different educational, provincial and racial classification that the greatest difference arise. The Kaplan-Meier survival estimates for these covariates also suggest significant divergence in the long-term equilibrium unemployment rates between subgroups according to these covariates.

Additionally clear divergence in the hazard ratios where found amongst the various demographic groups when evaluating the Aalen linear hazard model.

Surprisingly my initial analysis suggested that a matric qualification will not offer much benefit in the South African labour market and may in fact be associated with worse outcomes than having no secondary education. However, these results are likely driven by perceptions amongst employers of the relatively poor quality of the education at non-white schools. The results also highlighted the importance of technical training such as that provided by the National Technical Colleges, which offered labour market benefits comparable to those of a University degree. Additionally, the willingness of those without completed secondary education to participate in the informal economy may give them a competitive edge in the job search market over matrics.

Estimates based on gender and rural-urban differences suggested that the differences in these categories are quite small. The former suggest a slight benefit to females in the labour market. Regardless of the sources of variation between genders and rural-urban groups, the estimated hazard rates for both groups converge over time, as do their estimated survival probabilities. A caveat to this result is that there is clear evidence that labour market experiences of rural and urban dwellers differ significantly between provinces. The driving factors of these differences has not been investigated here but are likely to include the industrial makeup of different provinces, and the extent of migrant labour inflows to different urban areas.

In economies where there are many job seekers with few available positions companies are likely to rank potential employees and only send job offers to the best. In their ranking model of unemployment Diamond and Blanchard (1990) show that there can exist divergence in exit rates from unemployment between highly ranked and poorly ranked groups. This hypothesis seems too be borne out in the divergence documented amongst educational subgroups in the Cox-PH model and Aalen linear hazards model. Ranking of unemployed workers in the South African labour force can therefore be thought of as a combination of the quality of educational attainment, and geographical location.

A policy prescription from these results will emphasise the importance of extending technical education to the unskilled and improving the quality of education in historically disadvantaged schools. How to go about implementing these reforms lies outside the scope of this thesis.

Three potential avenues for future research raised by this thesis include:

Analysing short term spells of unemployment.

Little is known about the drivers of unemployment duration in the short term. This area of research is likely to be vital for assessing the impact of labour market interventions in returning retrenched workers to the labour force, and easing the assimilation of youths into the labour force. As no such data set currently exists it would have to be compiled via a frequently repeated survey with a 2 to 4 week gap between waves.

A comparative analysis of labour market conditions as decomposed by the urban and rural areas of all the provinces in South Africa.

As stated earlier large differences exist in labour market conditions based on a combination of provincial and rural-urban location. Special focus should be given the role played by urbanization trends, including intra-provincial urbanization.

Determining the drivers of labour absorption in the informal economy.

Some evidence has been given that the labour absorption in the informal economy may be higher than that of the formal economy. Future studies should attempt to assess to what extent this is being driven by wage differentials as opposed to restrictive labour market legislation. The findings of such a study are likely to have important consequences for potential labour market interventions, such as the proposed youth labour subsidy.

References

Banarjee A., Galiani, S., Levinsohn J., McLaren Z. & Woolard, I. 2008. Why has Unemployment Risen in the New SouthAfrica? *Economics of Transition* 16(4): 715-740.

Blanchard, J. B. & Diamond, P. 1990.*Ranking, Unemployment Duration and Wages.* NBER Working Paper 3387.

Breslow, N.E. 1974. Covariance Analysis of Censored Survival Data. *Biometrics*, 30, 89- 99.

Cichello, P. L., Fields, G. S. & Leibbrandt, M. 2005. Earnings and Employment Dynamics for Africans in Post-Apartheid South Africa: A Panel Study of KwaZulu-Natal. *Articles and Chapters.* Paper 263. Ithaca, New York: Cornell University.

Cleves M. A., Gould W. W. & Gutierrez R. G. 2004. *Introduction to Survival Analysis Using STATA: Revised Edition.* College Station: Stata Press.

Dinkelman and Ranchod. 2008. *Labour Market Transitions in South Africa: What can I learn from the matched Labour Force Survey?* SALDRU Working Paper 14.

Fourie, F. 2011. *The South African unemployment debate: three worlds, three discourses?* SALDRU Working Paper 63.

Grambsch, P.M. and Therneau, T.M. 1994. Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika* 81 (3), 515-526.

Hosmer D. & Lemeshow S. 1999. *Applied Survival Analysis: Regression Modeling of Time to Event Data.* New York: John Wiley and Sons.

Hosmer. D.W., and Royston, P. 2002. Using Aalen´s linear hazard model to investigate time-varying effects in the proportional hazards regression model. *The Stata Journal* 2(4), 331-350.

Kalbfleisch, J.D. and Prentice, R.L.1981. Estimation of the average hazard ratio. *Biometrika*, 68, 105-112.

Kingdon, G.G. & Knight, J. 2004. Unemployment in South Africa: the nature of the beast. *World Development,* 32(3): 391-408.

Klasen S. & Woolard I. 1999. Surviving Unemployment without State Support: Unemployment and Household Formation in South Africa. *Biennial Conference of the Economic Society of South Africa, 6-7 September. Youth unemployment.*

Klein J. P. & Moeschberger M. L. 2003. *Survivor Analysis: Techniques for Censored and Truncated Data: 2nd Edition.* New York: Springer.

McCall, J.J. 1970. Economics of information an Job Search. *Quarterly Journal of Economics* 84: 113-126.

Mortenson D.T. 1986. Job Search and Labor Market Analysis. *Ch. 15 Handbook of Labor Economics Vol. 2 (Editors Ashenfelter O. & Layard R.).* North-Holland: Amsterdam.

Mortenson, D. T. 1970. A Theory of Wage Employment Dynamics. *Microeconomic economic foundations of employment and inflation theory.(Editor E.S. Phelps et al)* New York: W. W. Norton.

Mortenson, D.T. Gronau, R. 1971. Information and Frictional Unemployment. *American Economic Review,* 61: 290-301.

Persson, I. 2002. *Essays on the Assumption of Proportional Hazards in Cox Regression.* [Available Online: http://www.google.com/url?sa=t&rct=j&q=essays%20on%20the%20assumption%20o f%20proportional%20hazards%20in%20cox%20regression&source=web&cd=1&ved=0CCIQFjAA&url=http%3A%2F%2Fuu.divaportal.org%2Fsmash%2Fget%2Fdiva2%3A161225%2FFULLTEXT01&ei=gF0lT7nPFYOn0QWfpInPCg&usg=AFQjCNEiqTWvEQ7QXu4dmWrW3SE5rn2vrA]

Schoenfeld, D. 1982. Partial residuals for the proportional hazard regression model. *Biometrika* 69:239-241.

Seeking, J. 2003. *Do South Africa's unemployment constitute an underclass?* Working

Stigler, G. J. 1961. The Economics of Information. *Journal of Political Economy,* 69: 213- 225.

Stigler, G. J. 1962. Information in the Labor Market. *Journal of Political Economy,* 70*:* 94- 104.

Statistics South Africa. 2006. Labour Force Survey Panel 2006: Beta Version.

Statistics South Africa. 2008. Labour Force Survey: Historical Revisions March Series 2000 to 2007.

Statistics South Africa. 2009. Labour Force Survey: Historical Revisions September Series 2000 to 2007.

Van den Berg. 2000. Duration Models: Specification, Identification, and Multiple Durations. *Handbook of Econometrics, Vol 5. (Editors: Heckman & Leamer).* North-Holland, Amsterdam.

Wittenberg M. 1999. *Job Search and Household Structure in an Era of Mass Unemployment: a Semi-parametric analysis of the South African Labour Market.* Working Paper 22. South African Network for Economic Research, University of Potchefstroom.

Statistical Appendix I: Technicalities of Survival Analysis

Note 1

Using the methodology of Greenwood (1926) we calculate an estimate of the variance of the Kaplan-Meier survivor estimate as:

$$Var\big(S(t)\big) = S^2(t) \sum_{j|t_j<t} \left(\frac{d_j}{n_j(n_j - d_j)}\right)$$

While the above function can be used for the estimation of standard errors I do not use it for the estimation of confidence intervals. Instead the estimation of confidence intervals I follow the methodology of Kalbfleisch and Prentice (2002) by first estimating the asymptotic variance of ln(-lnS(t)):

$$\theta^2(t) = \frac{\sum_{j|t_j<t}\left(\frac{d_j}{n_j(n_j - d_j)}\right)}{\left(\sum_{j|t_j<t} \ln\left(\frac{n_j - d_j}{n_j}\right)\right)^2}$$

Using theta as estimated above I can estimate a (1-alpha)% confidence interval of the survival function estimate as:

$$\hat{S}(t) exp[\pm z_{\frac{\alpha}{2}} \theta^2(t)]$$

Where

$$z \sim N(0,1)$$

Note 2

According the Cleves *et al* (2004) the estimation of the hazard function starts by determining the estimated hazard contribution, based upon the N-A cumulative hazard estimator:

$$\Delta H(t) = H\big(t_j\big) - H\big(t_{j-1}\big)$$

having established the estimated hazard contribution I estimate h(t) as

$$h(t) = b^{-1} \sum_{j=1}^{D} K\left(\frac{t - t_j}{b}\right) \Delta H(t)$$

Note 3

To obtain a confidence interval of the N-A cumulative hazard estimator I follow the methodology of Aalen (1978) by first estimating the estimation of the variance of H(t) as:

$$Var(H(t)) = \sum_{j|t_j<t} \frac{d_j}{n_j^2}$$

We can now estimate the confidence interval of H(t) as:

$$\hat{H}(t) exp[\pm z_{\frac{\alpha}{2}} \theta^2(t)]$$

Where I define

$$\theta^2(t) = \frac{Var(H(t))}{\big(H(t)\big)^2}$$

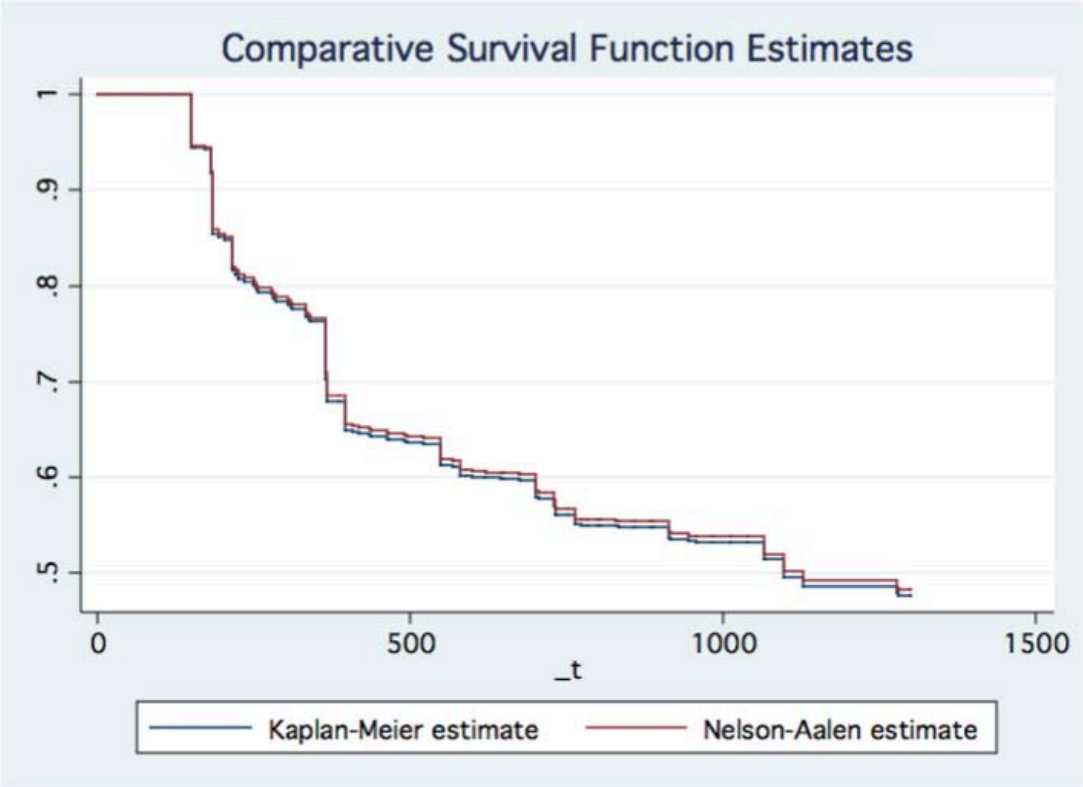
and z is defined as before.

Note 4

At this point it is worth clarifying that either the Kaplan-Meier or the Nelson-Aalen estimators can be used to estimate both the Survival function and the cumulative hazard function via the relationship:

$$S(t) = \exp[-H(t)] \Rightarrow H(t) = -\log S(t)$$

In fact the estimates for the K_M estimator and the Nelson-Aalen estimators will be asymptotically equivalent. However the small sample properties of the K-M estimator are superior for the estimations of the survivor function. Similarly the N-A estimator has better small sample properties for the estimation of the cumulative hazard (Klein and Moeschberger, 2003).



A graphical comparison of the estimated survival functions is given above. Note that the N-A estimator is less than the K-M estimator of the survival function for all values of t, similarly the K-M estimator of the cumulative Hazard function is less than the corresponding N-A estimator. This is to be expected as

was proven using a Taylor expansion by Hosmer and Lemeshow (1999).

Note 5

The Wilcoxon test´s differences in weighting leave it susceptible to unreliability if censoring patterns differ significantly between subsamples. Note however that the weighting of the Wilcoxon test are higher for early failure times, the test should therefore continue to indicate statistical differences between subgroups assuming that initial hazard rates differ (Cleves *et al.* 2004).

Note 6

The assumed multiplicative nature of the hazard function means that the underlying distributional qualities of the baseline hazard are not necessary for estimation. Additionally the Cox-PH model should contain no constant as this would be included in the baseline hazard. Using the functional form of the hazard function under the Cox PH model I can obtain the relate the relative relationship of risk of failure between individuals as:

$$\frac{h(t|\boldsymbol{x_j})}{h(t|\boldsymbol{x_m})} = \frac{\exp(\boldsymbol{x_j\beta_x})}{\exp(\boldsymbol{x_m\beta_x})}$$

This relation is used for the maximum likelihood estimation of the coefficients. Consider for example the simplified case of 4 observed failures with a single covariate and distinct failure times. Defining P_i as the conditional probability of failure for individual i at time of failure, I define the likelihood function as:

$$L(\beta) = \prod_{i=1}^{4} P_i$$

which can be rewritten as

$$L(\beta) = \prod_{i=1}^{4} \frac{\exp(x_i\beta)}{\sum_{j\in R_i}\exp(x_j\beta)}$$

From this point maximum likelihood estimation can be used to estimate beta.

Complicating the calculation in my case, is the existence of multiple tied observations, or observations where failure is recorded at the same point in time for multiple observations. While multiple methods exist for dealing with tied failures, the default used by STATA is the Breslow (1974) approximation as it is the least computationally intensive (Cleves et al. 2004). Returning to the example given above, suppose that instead of distinct failure times observation 2 and 3 were recorded as failing at the same point in time. Now defining P_{ij} as the probability that individual I fails before individual j, the Breslow method estimates

$$P_{23} = P_{32} = \prod_{i=2}^{3} \frac{\exp(x_i\beta)}{\sum_{j\in R_i}\exp(x_j\beta)}$$

and the contribution of the tied observations to the likelihood function is estimate as:

$$P_{23} + P_{32}$$

Note 7

Following the definition given by Kalbfleisch and Prentice (1981), the geometric average of the hazard ration in the case of a binomial covariate is:

$$\theta(W) = -\int_0^{\infty} \frac{h(t|x=1)}{h(t|x=0)} dW$$

where

$$W = S(t|x=1)^{1/2} S(t|x=0)^{1/2}$$

The difference between the estimated hazard ratio and the geometric average of the hazard will be increase where hazard rates are increasing or decreasing to a greater extent. It will also differ more with higher levels of censoring or smaller sample sizes (Persson, 2002).

Note 8

Estimation of the Aalen linear hazards model makes use of the estimated cumulative hazard. It uses the same assumptions as traditional OLS estimation where the estimates of the regression coefficients at time j:

$$\widehat{\boldsymbol{\beta}_x}(j) = \left(\boldsymbol{X_j}'\boldsymbol{X_j}\right)^{-1}\boldsymbol{X_j}'\boldsymbol{y_j}$$

Where **y** is a nx1 vector of recorded survival times. Accordingly the estimator of the vector of the cumulative regression coefficients is

$$\widehat{\boldsymbol{B}}(t) = \sum_{\boldsymbol{t_j} \le \boldsymbol{t}} \widehat{\boldsymbol{\beta}}_{\boldsymbol{x}}(\boldsymbol{t_j})$$

The estimated cumulative hazard estimate as a function of time and the covariates can then be described as:

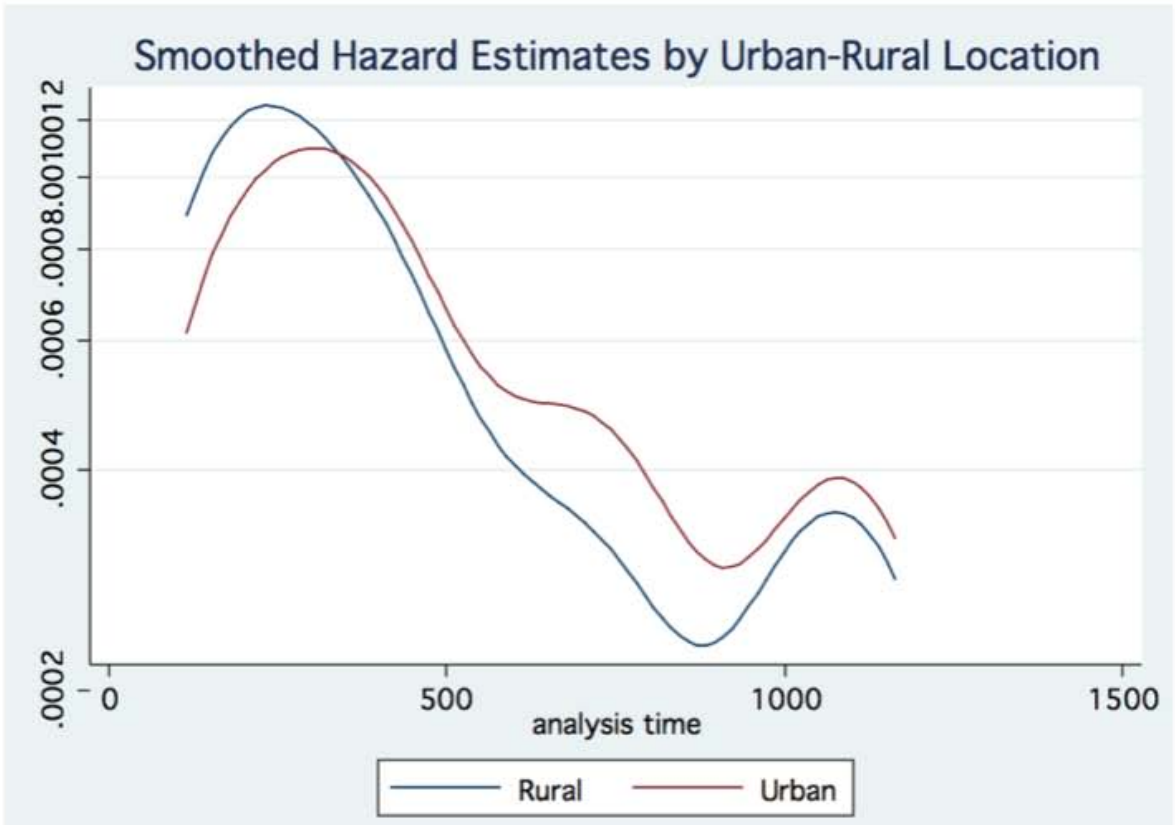
$$H\left(t, \boldsymbol{X_i}, \widehat{\boldsymbol{B}}(t)\right) = \widehat{\boldsymbol{B}}(t)\boldsymbol{X_i}$$

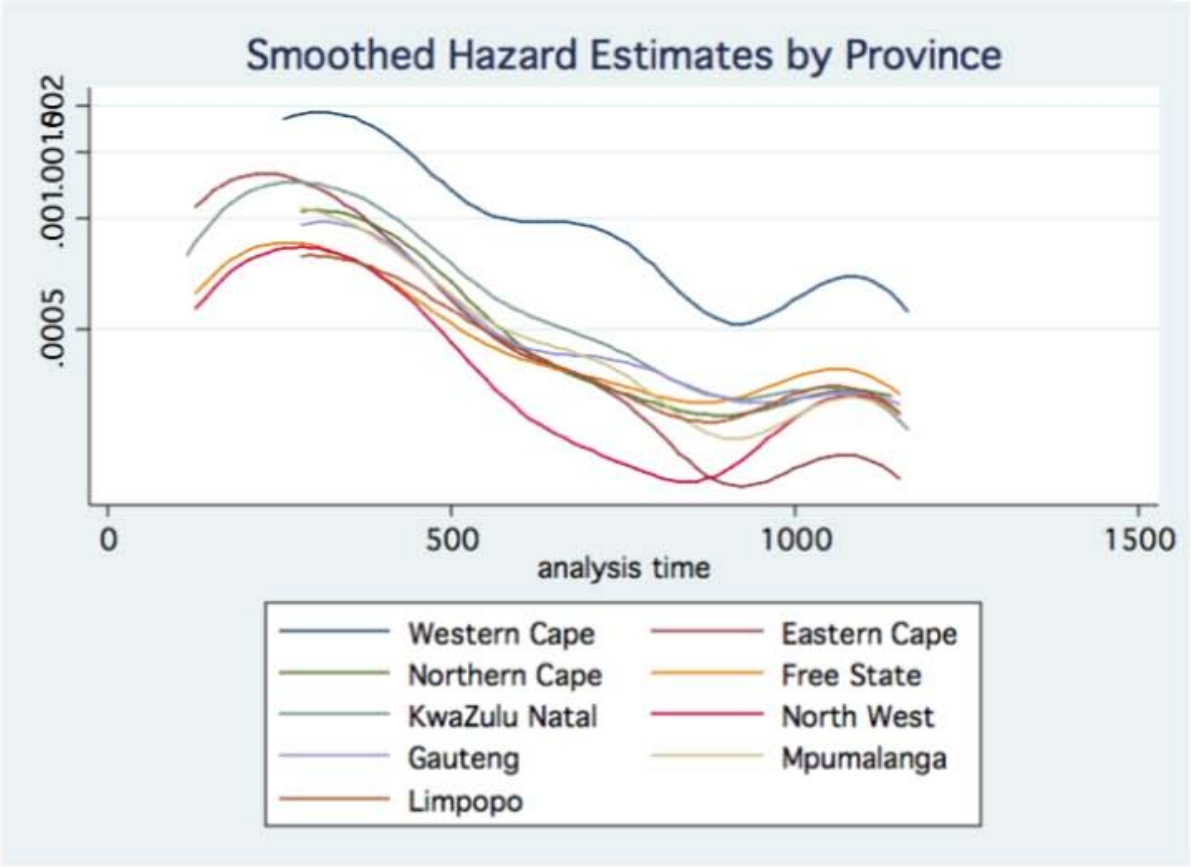
Allowing for estimation via OLS estimation. The Aalen linear hazard model repeatedly estimates the cumulative regression coefficients using the cumulative hazard estimations. Deviations over time in the cumulative regression coefficient are used to estimate the Aalen model. As the Aalen model uses multiple regressions to estimate the coefficients it is quite computationally intensive.

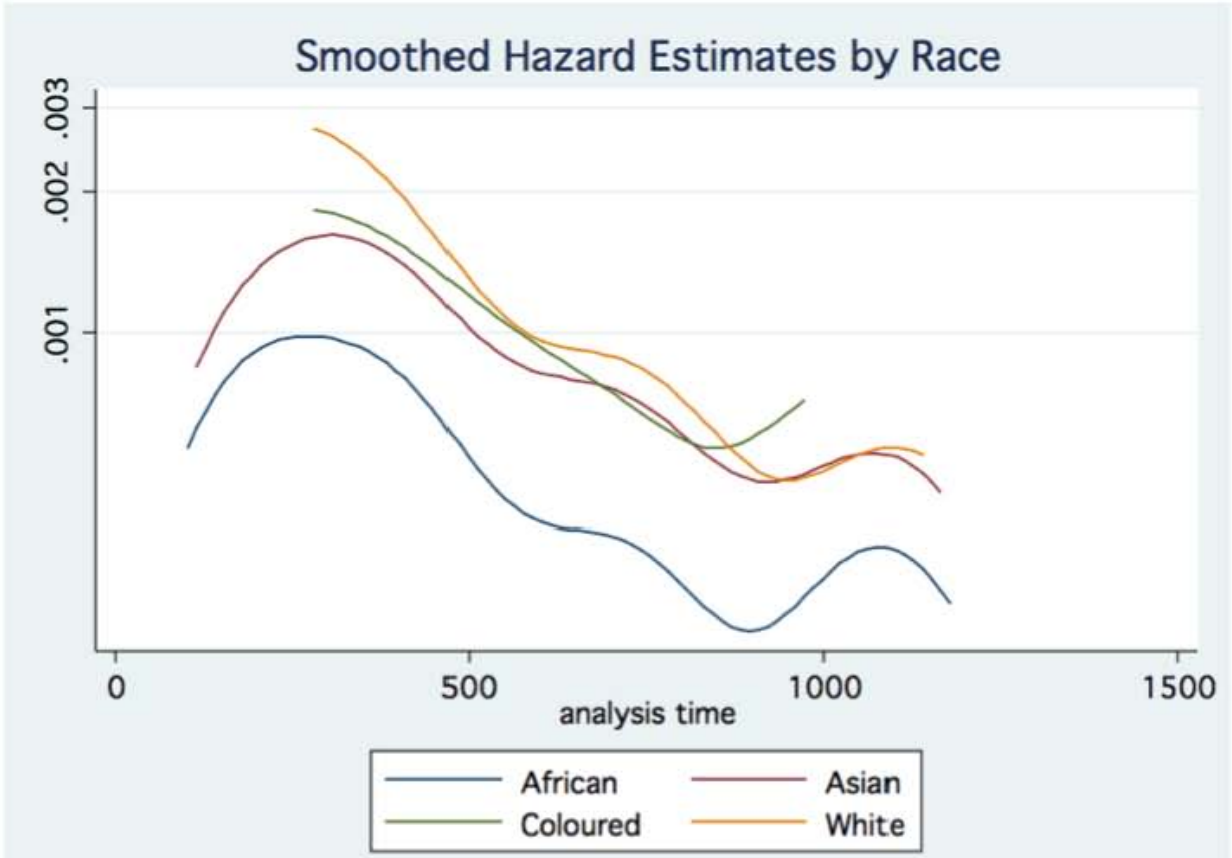
Statistical Appendix II: Extended Statistical Analysis

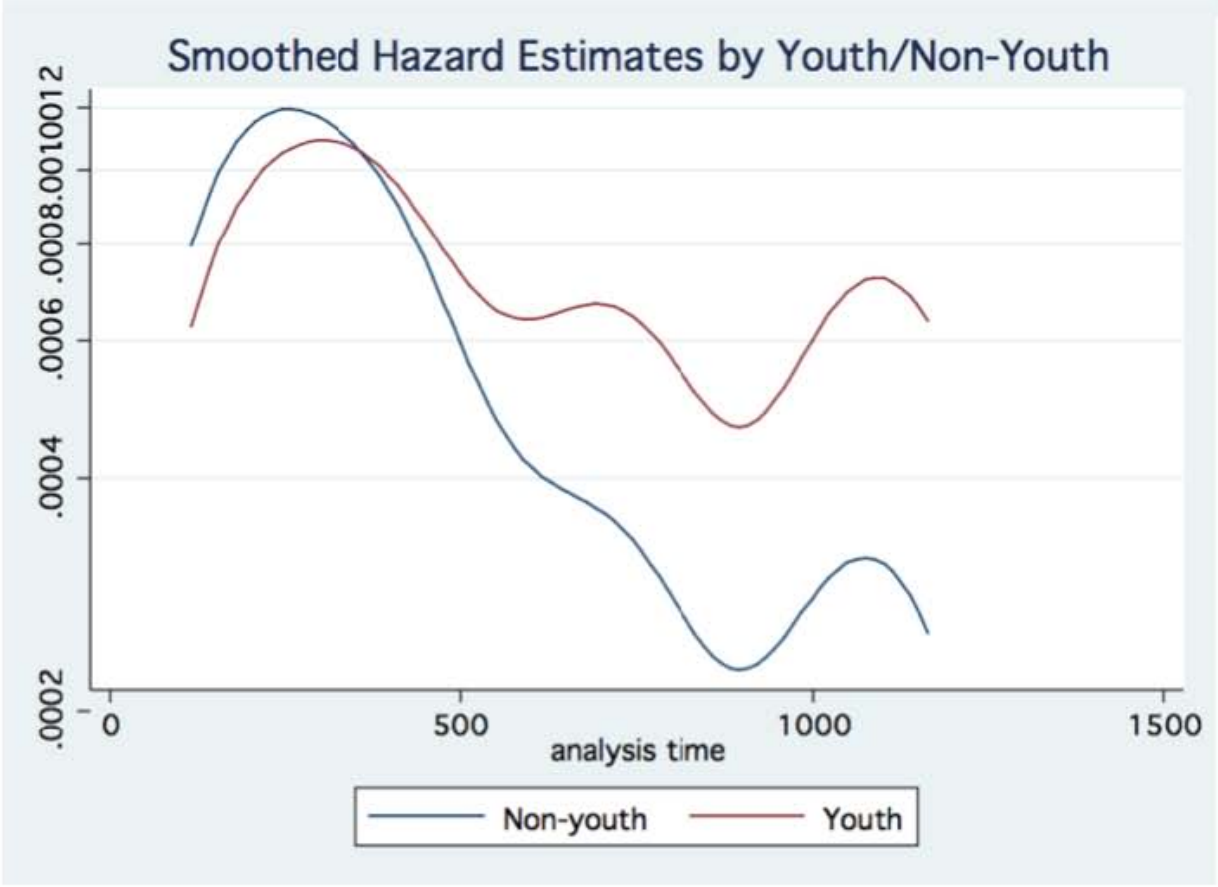
Section 3.2 Individual Covariate Analysis

Graphical Representations of the Smoothed Hazard Functions









3.2.1 Gender Based Differences

In the years following the transition to democracy there has been significant growth in labour force

participation amongst women in South Africa, especially amongst unskilled African women (see Banarjee *et al.* 2007). I may therefore expect that jobs traditionally filled by women would be more easy to fill extending average female unemployment spells.

Table 2.2.1.1 Survivor functions by Gender

Gender	Male	Female
Analysis Time		
1	0.9996	0.9996
217	0.8397	0.7972
434	0.6576	0.6382
867	0.5513	0.5468
1300	0.4801	0.4747

Table 2.2.4.2 Testing for Statistical differences in the Survivor Function by Gender

Stratification	None		Education		Urban	
	Test Statistic*	p-value	Test Statistic*	p-value	Test Statistic*	p-value
Log-Rank	6.69	0.01	5.93	0.01	5.25	0.02
Wilcoxon	21.77	0.00	21.71	0.00	131.22	0.00
Stratification	**Race**		**Education, Urban & Race**			
	Test Statistic*	p-value	Test Statistic*	p-value		
Log-Rank	7.43	0.01	3.39	0.07		
Wilcoxon	11.65	0.00	20.30	0.00		

The estimated hazard function suggests that that the relationship between gender and unemployment duration is mixed. While men initially have higher hazard rates of transitioning out of unemployment after losing their jobs, this trend is reversed later. This reversal in the hazard rates women continue to display lower estimates survivor functions for all periods considered. The estimated differences in the survival function between the sexes is, however, relatively small. By 1300 days of analysis time the difference in survival probability to only 0.54 percentage points indicating male and female survival rates converge over time.

2.2.2 Education based differences

The importance of education in the labour market has been stated, *ad infinitum,* by labour economists. Within the search model there education can clearly effect both the arrival rate of job offers and the quality of offers received thereby affecting the probability of transitioning out of unemployment.

Table 2.2.2.1 Survivor functions by Educational Attainment

Education Level	No-High school	Some High School	Matric	NTC
Analysis Time				
1	0.9998	0.9996	0.9992	1
217	0.7536	0.8539	0.8472	0.8474
434	0.6104	0.689	0.6653	0.5599
867	0.5477	0.5892	0.5342	0.3593
1300	0.4919	0.5216	0.444	0.2005

Education Level	Diploma/Certificate	Degree/PostGraduate Degree
Analysis Time		
1	1	1
217	0.8419	0.6738
434	0.6489	0.4467
867	0.5484	0.33
1300	0.4357	0.2437

Table 2.2.4.2 Testing for Statistical differences in the Survivor Function by Educational Attainment

Stratification	None		Gender		Urban	
	Test Statistic*	p-value	Test Statistic*	p-value	Test Statistic*	p-value
Log-Rank	148.69	0.00	150.78	0.00	153.35	0.00
Wilcoxon	153.61	0.00	159.48	0.00	131.22	0.00

Stratification	Race		Gender, Urban an Race	
	Test Statistic*	p-value	Test Statistic*	p-value
Log-Rank	146.46	0.00	72.01	0.00
Wilcoxon	194.64	0.00	116.26	0.00

As expected, university graduates show the greatest initial labour market absorption. Surprisingly workers with matric show lower hazard rates than those classified as having attended no high-school. Another surprise is the shape of the hazard function amongst NTC graduates, notably it does not steadily decline, but rather, shows a pronounced u shape over time. It should be remembered, however, that NTC graduates include all 3 levels of NTC qualifications and therefore represent the most heterogenous group amongst all education groups. The NTC qualification are primarily forms of technical training, the duration data therefore re-emphasises the importance of technical skills in the South African labour market.

The estimated survivor function by educational level suggest that there is no convergence in labour absorption between educational subgroups. In the sample NTC graduates were most likely to

transition into employment after a period of unemployment while those who completed only some years of high schooled faired the worst.

2.2.3 The Rural-Urban Divide

The residual effects of apartheid era policies on the geographical distribution of labour has also been listed as a potential source of high unemployment levels in South Africa (Banarjee *et al.* 2007). The derived urban-rural indicator variable offers an excellent opportunity to test for the effects of geographical location on the duration of unemployment spells.

Table 2.2.3.1 Survivor functions by Urban-Rural Status

Gender	Rural	Urban
Analysis Time		
1	0.9996	0.9996
217	0.7673	0.8492
434	0.6192	0.6659
867	0.5416	0.5548
1300	0.4754	0.4796

Table 2.2.3.2 Testing for Statistical differences in the Survivor Function by Urban-Rural Status

Stratification	None		Education		Gender	
	Test Statistic*	p-value	Test Statistic*	p-value	Test Statistic*	p-value
Log-Rank	24.79	0.00	19.68	0.00	23.30	0.00
Wilcoxon	73.15	0.00	68.81	0.00	75.01	0.00
Stratification	**Province**		**Education, Gender and Province**			
	Test Statistic*	p-value	Test Statistic*	p-value		
Log-Rank	58.89	0.00	32.39	0.00		
Wilcoxon	100.35	0.00	56.92	0.00		

The estimated survivor function suggests that rural workers are initially absorbed faster into the labour force, before survival rates between rural and urban workers eventually converge.

2.2.4 Racial Differences

Racial differences in labour market experience are generally accepted as residual effects of apartheid era policies. Under these policies Whites received preferential access to the labour market, social services, higher quality education and did not face restrictions on the restrictions on mobility impose on Africans by the pass system. The other two racial categories faced less odious restrictions than the Africans while still not gaining access to the privileges afforded to Whites (Lam *et al.* 2009).

Table 2.2.4.1 Survivor functions by Gender by Racial Group

Race	African	Coloured	Asian	White
Analysis Time				
1	0.9996	0.9994	1	1
217	0.8401	0.809	0.6949	0.5031
434	0.6866	0.5641	0.4735	0.3065
867	0.5973	0.4172	0.3521	0.218
1300	0.5259	0.3328	0.2654	0.1761

Table 2.2.4.2 Testing for Statistical differences in the Survivor Function by Racial Group

Stratification	None		Education		Gender	
	Test Statistic*	p-value	Test Statistic*	p-value	Test Statistic*	p-value
Log-Rank	551.01	0.00	572.89	0.00	552.41	0.00
Wilcoxon	485.8	0.00	381.56	0.00	462.57	0.00
Stratification	**Urban**		**Education, Gender and Urban**			
	Test Statistic*	p-value	Test Statistic*	p-value		
Log-Rank	694.83	0.00	444.81	0.00		
Wilcoxon	694.22	0.00	377.20	0.00		

As expected, labour absorption in the sample is greatest amongst Whites and least amongst Africans. These estimates do not show any signs of convergence; and at the time of censoring survival probability amongst s is 34.98 percentage points less than that of Africans.

2.2.5 The unemployed youth

The issue of youth unemployment has attracted considerable attention internationally. In South Africa

high and stable levels of youth unemployment have raised concerns over the long-term social consequences of a "lost generation". The problem of youth unemployment in South Africa is closely linked to poor educational outcomes (Lam *et al.* 2009). For the purposes of my analysis I will define a youth as any individual who was less than 30 years old for the duration of the period that she was in observed in my data set.

Table 2.2.5.1 Survivor functions by Youth-Old

	Non-youth	Youth
Analysis Time		
1	0.9997	0.9994
217	0.7997	0.8412
434	0.634	0.6651
867	0.5512	0.52
1300	0.4913	0.3926

Table 2.2.5.2 Testing for Statistical differences in the Survivor Function by Youth-Old

Stratification	**None**		**Education**		**Gender**	
	Test Statistic*	**p-value**	**Test Statistic***	**p-value**	**Test Statistic***	**p-value**
Log-Rank	0.32	0.57	0.05	0.82	0.14	0.70
Wilcoxon	20.19	0.00	1.16	0.28	24.00	0.00
Stratification	**Urban**		**Education, Gender and Urban**			
	Test Statistic*	**p-value**	**Test Statistic***	**p-value**		
Log-Rank	0.37	0.54	0.01	0.92		
Wilcoxon	11.86	0.00	3.04	0.08		

The estimated hazard function supports the idea that the youth initially face a disadvantage in the labour market, as might be expected due the relative lack of experience amongst young workers (Figure 2.2.5.1). However, this trend is reversed over time and at the time of censoring youths have an estimated survival probability nearly 10% lower than older workers (Table 2.2.5.1). Results from my youth indicator variable therefore support the hypothesis that younger workers may take more time than older workers to find a suitable job match.

2.2.6 Provincial differences

South Africa´s nine provinces differ significantly with regards to levels of economic development and types of industry. It seems entirely plausible that these factors would influence labour demand, and thereby, the job search model.

Table 2.2.6.1 Survivor functions by Province

	Western Cape	Eastern Cape	Northern Cape	Free State	KwaZulu Natal
Analysis Time					
1	1	0.9989	1	0.9991	0.9995
217	0.7754	0.7333	0.8569	0.8462	0.789
434	0.5026	0.5896	0.6601	0.7124	0.6168
867	0.3507	0.5133	0.5689	0.6152	0.5105
1300	0.2725	0.4714	0.4877	0.5269	0.4425
	North West	**Gauteng**	**Mpumalanga**	**Limpopo**	
Analysis Time					
1	0.9992	1	1	1	
217	0.8625	0.8663	0.8007	0.864	
434	0.7185	0.687	0.6425	0.7316	
867	0.6555	0.589	0.5477	0.6282	
1300	0.5781	0.5054	0.4877	0.5457	

Table 2.2.6.2 Testing for Statistical differences in the Survivor Function by Province

Stratification	**None**		**Education**		**Gender**	
	Test Statistic*	**p-value**	**Test Statistic***	**p-value**	**Test Statistic***	**p-value**
Log-Rank	265.03	0.00	271.49	0.00	264.98	0.70
Wilcoxon	256.35	0.00	239.66	0.00	266.77	0.00
Stratification	**Urban**		**Education, Gender and Urban**			
	Test Statistic*	**p-value**	**Test Statistic***	**p-value**		
Log-Rank	300.10	0.00	237.47	0.00		
Wilcoxon	227.80	0.00	187.55	0.00		

The estimated hazard functions suggest that, while some variation exists between provinces, only

the Western Cape province shows remarkable difference in labour absorption, at the provincial level. This higher level of labour absorption is estimated as lasting for the entire range of analysis and is matched by a significantly lower survival probability at the time of censorship, with a 27.25% probability of survival at 1300 days. Of the other provinces North West stands out as having the worst labour absorption amongst the unemployed with a 57.81% probability of survival at the time of censoring. All other provinces had final survival probabilities between that of North West and Kwa- Zulu Natal.

Section 3.3.3 : Cox PH model analysis

3.3.3.1 Gender Based Differences

The gender indicator variable suggests hat female participants had a hazard rate about 8.2% greater than that of men. The Cox diagnostic test rejects the proportionality assumption, suggesting convergence in hazard rates between genders over time. The Grambsch-Therneau diagnostic confirm the Cox diagnostic test in rejecting the proportionality assumption.

When controlling for other covariates the estimated gender differences in hazard rates between men and women increases to 16.9%, the results of both diagnostic tests continue to reject the proportionality assumption.

3.3.3.2 Educational Differences

The education variable suggest that, as expected, all other education groups had lower hazard rates than university graduates. Note however, that the difference between University graduates and NTC graduates was not statistically significant. The Cox diagnostic test roundly rejects the proportionality assumption and lends support for convergence between the hazard rates of university graduates and all other groups. The difference in the estimated hazard ratios between educational levels reduce in magnitude once I control for other covariates. The coefficient estimated for NTC graduates suggest higher hazard rates amongst NTC graduates than University graduates, this coefficient is, however, not statistically significant.

When controlling for the other covariates the Cox diagnostic test rejects proportionality only for the no-high school and Diploma variable. The Grambsch-Therneau diagnostic test additionally rejects proportionality for the Matric and Some high school indicator variables. The Cox diagnostic test now predicts divergence in the hazard rates between those with University degrees and those without.

3.3.3.3 Rural Urban Differences

The estimated hazard ratio between rural and urban residents suggest that the hazard rate of Urban dwellers was approximately 85.86% that of rural dwellers. The Cox diagnostic test reject proportionality and suggest some convergence between hazard ratios, the Grambsch-Therneu diagnostic test also rejects the proportionality assumption. Controlling for all variables the estimated difference in hazard ratios between urban and rural dwellers increases, both diagnostic tests continue to reject the proportionality.

3.3.3.4 Provincial Differences

Provincial hazard ratios suggest that all other provinces have statistically lower hazard rates on aggregate than the Western Cape province. The Cox diagnostic test rejects proportionality only for the Eastern Cape, Kwa-Zulu Natal and North West and Mpumulanga. For all of these the Cox-PH model suggests divergence of hazard rates. The Grambsch-Therneu diagnostic test concurs with the findings of the Cox diagnostic test.

Once I control for the other covariates the estimates on the provincial covariates change significantly, with the differences in hazard rates between the Western Cape and six of the eight provinces increasing. This may indicate that provincial factors have a larger than expected effect on the job search function once I control for other factors. The Cox diagnostic test now does not reject proportionality for all the provinces apart from Limpopo and the Eastern Cape, suggesting divergence in both cases. The Grambsch-Therneu test additionally rejects for proportionality for the Free-State province indicator variable.

3.3.3.5 Racial Differences

The estimated Hazard Ratios confirm that all other racial grouping have higher hazard estimates than Africans. The Cox diagnostic test does not reject proportionality between Coloureds and Africans. For Asians and Whites, the Cox tests respectively suggests divergence and convergence relative to the African baseline. The Grambsch-Therneu test results concur with the Cox test for Asians and Coloured but does not reject for proportionality between Whites and Africans.

Controlling for other variables increases the estimated difference in hazard rates between Africans and Whites and Africans and Asians respectively while increasing it between Africans and Coloureds. The Cox diagnostic test rejects proportionality only for the White indicator variable, suggesting convergence over time while the Grambsch- Therneu test additionally rejects proportionality for the

Coloured indicator variable.

3.3.3.6 Age based Difference

The estimates on the age variables suggest a positive but diminish effect of age on the hazard ratio. This effect is most likely linked to the accumulation of labour market experience with age. Once I control for this relationship my youth variable has large positive effect on the hazard rates. Both diagnostic test reject proportionality.

Once I control for other covariates the coefficient estimates for the age variables remain largely the same while the coefficients estimates of the youth indicator variable decrease somewhat. The Cox-diagnostic test does not reject the proportionality assumption while the Grambshc Therneau diagnostic test delivers mixed results.

3.3.3.6 Unemployment quintiles

Table 3.3.3.6 Estimates of Unemloyment Quintiles in the Cox PH Model As expected, university graduates show the greatest initial labour market absorption.

	Cox PH Model Hazard Ratio estimate			**Cox diagnostic test**		**Grambsch and Therneau diagnostic test**		
	Hazard ratio	Std Error	p>|z|	Beta xt	p>|z|	chi sq	p>|z|	DF
Measured independently						28.16	0.000	4
First	1.390	0.072	0.000	-0.00074	0.000	16.14	0.000	1
Second	1.754	0.090	0.000	-0.00021	0.217	0.05	0.822	1
Third	1.458	0.079	0.000	-0.00077	0.000	7.37	0.007	1
Fifth	1.360	0.077	0.000	-0.00033	0.073	4.17	0.041	1
Controling for other Covariates								
First	1.002	0.092	0.984	0.00009	0.785	0.04	0.833	1
Second	2.510	0.188	0.000	-0.00027	0.371	1.85	0.174	1
Third	1.218	0.090	0.008	0.00024	0.384	0.74	0.389	1
Fifth	1.597	0.135	0.000	0.00164	0.000	15.83	0.000	1

The relationship between the provincial unemployment rate and the hazard rate was clearly non-linear, with the 4th quintile showing the lowest hazard rates. For this reason the unemployment control variable was included via dummy variables for each quintile. The diagnostics test give mixed

reports on the proportionality of the unemployment quintile variable.

Controlling for other covariates the unemployment quintile variable finds statistical significant coefficients for all indicators except the first quintile´s estimate. Both the Cox and the Grambsch-Therneau diagnostic tests do not reject proportionality except in the case of the fifth quintile indicator variable.