Assessing the patterns of selection into high school dropout and high school graduation: Evidence from South Africa *

Godstime Eigbiremolen †

Job market paper

Abstract

This paper assesses the patterns of selection into high school dropout and high school graduation using longitudinal data from South Africa. Employing a competing risks model, I estimate a reduced-form measure of self-selection and establish to what extent selection matters. Results suggest that individuals who choose to drop out of high school are not systematically different from those who complete high school education in terms of unobservables. Rather, the factors that drive selection into dropout and graduation are observable individual and family characteristics. This runs contrary to typical findings in developed countries, which suggest that high school dropouts are individuals with relatively low ability, low expectations, and a set of negative preferences.

JEL classification: I21, A21, I25, I28

Key words: high school, dropout, graduation, unobserved heterogeneity, competing risks

*I am indebted to Andrey Launov for very useful feedback, comments and encouragement. I am also grateful to seminar participants at the University of Kent, Canterbury and the University of Sussex, Brighton for helpful comments and suggestions. As always, all remaining errors are mine.

†School of Economics, University of Kent: ge73@kent.ac.uk
1 Introduction

Dropping out of high school has both economic and social costs. Individuals who drop out may have difficulties finding good jobs due to low human capital accumulation that persists into the future in the form of low wages. Limited work opportunities and low earnings in the labour market may lead to inter-generational cycle of poverty. Furthermore, dropouts may not be able to fully contribute to national economic growth and development, and may also be susceptible to social vices (Hill, 1979; Cawley, Heckman, & Vytlacil, 2001; Lochner & Moretti, 2004).

The social and economic costs associated with high school dropout are particularly severe in South Africa where high school dropout rate has reached a point of national crisis. Approximately 60% of first graders will ultimately drop out rather than complete 12th Grade. Likewise, by Grade 12, only 52% of the age appropriate population remain enrolled (Weybright, Caldwell, Xie, Wegner, & Smith, 2017). As a result of low human capital accumulation, many young people who drop out remain unemployed for a long time. South Africa currently has the highest rate of youth unemployment (percent of total labor force ages 15-24) in the world (ILO, 2019).

The first step in addressing the problem of high school dropout is to identify possible factors that drive selection into dropout and graduation. There is an understanding in the literature that apart from observable characteristics, individuals may choose to drop out or complete high school education due to unobserved factors. Individuals may vary in motivation, innate ability, preferences and expectations. Thus, the patterns of dropout and graduation that we observe may be driven largely by unobservable characteristics. This knowledge is important to guide policies aimed at addressing the problem of high school dropout. For example, if those who choose to drop out from high school are systematically different from those who graduates, incentives policies may not be enough to keep them in school.

In this paper, I adopt a flexible framework that allows me to quantify selection on unobservables between dropouts and graduates and establish to what extent this selection matters. To do this, I write down a flexible discrete-
time conditional competing risks model that allows for general correlated risks across destinations. I treat individual differences that may exist between dropouts and graduates as both observed and unobserved. I then use the structure of my model to estimate a reduced-form measure of selection into dropout and graduation.

As a reduced-form measure of selection on unobservables, I propose to consider the correlation coefficient of the bivariate distribution of unobserved heterogeneity across two possible exit states: graduation and dropout. A statistically significant negative correlation coefficient means that an individual self-selects into high school completion and out of dropout or they self-select into dropout and out of graduation. Those who choose more years of schooling due to the fact that they are high ability have higher return in the labour market (Carneiro, Heckman, & Vytlacil, 2011). Such individuals are “good” in one dimension (i.e., either dropout or graduation), and “bad” in the other dimension (i.e., either dropout or graduation) at a given time.

For example, if an individual knows that there is a positive return to high ability (e.g., high wages in the labour market) and they are high-ability type; they will self-select into education (high school completion) and self-select out of dropout. Here, education is considered as “good” and there is a common belief that the return to education is positive (Dominitz & Manski, 1996). On the other hand, even though a low-ability person knows about the positive returns to human capital, they will not want to continue in education because it is more difficult for them to go through the curriculum. They will rather drop out fast and explore other possibilities. This is consistent with the theory of return to human capital (Lindsay, 1971; Goldin, 1999).

A statistically significant positive correlation coefficient means that an individuals self-select into education, just as they self-select into dropout. That is, they can do very well as graduates, just as they can do very well

---

1The method I proposed to assess the pattern of self-selection into high school dropout and high school graduation is similar to the Heckman selection model in principle. The bivariate distribution of unobserved heterogeneity in my approach is not fundamentally different from the bivariate distribution of errors in the Heckman model. The main difference is that while the model I used has a discrete bivariate distribution with 2 points of support, Heckman model has a continuous bivariate normal distribution.
as dropouts. For example, a student who has the ability to graduate from high school may choose to drop out because they do not believe that extra years in education will add to their productivity. This is particularly true for South Africa where the quality of education is relatively poor, especially for blacks (Case & Yogo, 1999; Selod & Zenou, 2003; van der Berg, 2007). This positive sorting suggests that individuals who select into graduation will have high lifetime earnings in jobs requiring high school diploma, while they will also earn high lifetime wages in jobs that do not require high school diploma if they had chosen to drop out of school (Willis & Rosen, 1979; Carneiro, Hansen, & Heckman, 2003).

My estimates however show no systematic pattern of self-selection into high school dropout and graduation in terms of unobservables. Rather, selection into dropout or graduation is driven by observable individual and family characteristics. This runs contrary to the widely held notion that compared to graduates, high school dropouts are individuals with relatively low innate ability, low motivation, low expectations, and a set of negative preferences.

There are studies in the education and self-selection literature that are close to my work, although my estimation approach is different from what exists in the literature. Roy (1951) provides one of the earliest discussions of self-selection, albeit in the context of occupational choices. Willis and Rosen (1979) extend Roy’s self-selection theoretical model and analyze schooling decision (high school versus college) in a non-hierarchical structure. However, the authors could not quantify the nature of self-selection because they were unable to identify the distributions of counterfactual choices in their models. In this paper, I am able to quantify the nature of self-selection and show to what extent selection matters.

Carneiro et al. (2011) extend Willis and Rosen’s self-selection model and also apply it to high school and college education choices. Their first extension is to account for the uncertainty in the returns to education. They also differentiate between present value income-maximizing and utility-maximizing evaluation of schooling choices. Unlike Willis and Rosen (1979), the authors were able to identify the distributions of counterfactual choices in their structural equation model and quantify the nature of selection. My
paper is most closely related to Carneiro et al. (2011) in overall objective (i.e., estimation of self-selection in schooling choices). What I do differently is to show an alternative estimation strategy and apply my analysis to the decision to drop out of high school or complete high school education.

The paper by Dolton and van der Klaauw (1999) is technically close to my estimation approach. The authors used a dependent competing risks framework that allows for a flexible semi-parametric specification of both the duration dependence and unobserved individual heterogeneity. While they analyze teacher’s decision to quit their job into either non-teaching career or non-working state, I employed a similar model to assess the pattern of self-selection into high dropout and graduation.

There is also a body of work in the literature that describes observable factors that determine high school dropout, conditional on unobserved heterogeneity. Some of these factors include compulsory school attendance (Angrist & Keueger, 1991); family characteristics (Cameron & Heckman, 2001; Oreopoulos & Page, 2006); test scores, cognitive and academic ability (Eckstein & Wolpin, 1999; Foley, Gallipoli, & Green, 2014); conditional cash transfer (Glewwe & Kassouf, 2012); motivation and expectations (Eckstein & Wolpin, 1999); university access (Bedard, 2001); and high school desegregation (Guryan, 2004). This paper also adds to these existing evidence in the literature, especially in a developing country context.

The rest of the paper is organized as follows. Section 2 describes the context and data; section 3 discusses the formulation of models and issues of identification; empirical results are discussed in section 4; while section 5 concludes and presents policy implications.

2 Data and context

This paper is based on the Cape Area Panel Study (CAPS). The data is a 5-wave longitudinal study of the lives of youths and young adults in metropolitan Cape Town, South Africa. The CAPS is the first data in Sub-Saharan African region designed to follow the lives of youths and young adults over a considerable time. The first wave of the survey randomly collected informa-
tion from young people aged 14-22 in August-December, 2002. Wave 1 also collected information on all members of these young people’s households, as well as a random sample of households that did not have members aged 14-22. A third of the youth sample was re-interviewed in 2003 (wave 2a) and the remaining two-thirds were re-visited in 2004 (wave 2b). The full youth sample was then re-interviewed in both 2005 (wave 3) and 2006 (wave 4). For wave 5, full face-to-face interviews was carried out in 2009 with the sample comprising all respondents interviewed in any of Waves 2a, 3 or 4.

Table 1: Descriptive statistics

<table>
<thead>
<tr>
<th>Variables</th>
<th>Dropouts</th>
<th></th>
<th>Graduates</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>St.Dev</td>
<td>Mean</td>
<td>St. Dev</td>
</tr>
<tr>
<td>Sex (male)</td>
<td>0.470</td>
<td>0.499</td>
<td>0.454</td>
<td>0.498</td>
</tr>
<tr>
<td>Local (urban)</td>
<td>0.787</td>
<td>0.409</td>
<td>0.811</td>
<td>0.391</td>
</tr>
<tr>
<td>Reservation wage (wage)</td>
<td>1726</td>
<td>2700</td>
<td>1406</td>
<td>1777</td>
</tr>
<tr>
<td>Time in household work</td>
<td>5.878</td>
<td>5.956</td>
<td>5.574</td>
<td>5.603</td>
</tr>
<tr>
<td>Test scores</td>
<td>8.330</td>
<td>5.558</td>
<td>10.735</td>
<td>5.532</td>
</tr>
<tr>
<td>Distance to school</td>
<td>19.729</td>
<td>14.590</td>
<td>19.773</td>
<td>14.833</td>
</tr>
<tr>
<td>Teacher absenteeism</td>
<td>0.142</td>
<td>0.350</td>
<td>0.106</td>
<td>0.308</td>
</tr>
<tr>
<td>Mother’s education</td>
<td>8.642</td>
<td>3.258</td>
<td>9.530</td>
<td>3.212</td>
</tr>
<tr>
<td>Mother helped with homework</td>
<td>0.200</td>
<td>0.399</td>
<td>0.224</td>
<td>0.417</td>
</tr>
<tr>
<td>Censored spell</td>
<td>37%</td>
<td></td>
<td>44%</td>
<td></td>
</tr>
<tr>
<td>Observations</td>
<td>1,539</td>
<td></td>
<td>1,314</td>
<td></td>
</tr>
</tbody>
</table>

Notes: Reservation wage is in South African Rand.

The sample for analysis compromises individuals that were enrolled in high school at the beginning of the survey (2002). These individuals are then followed over time until they either dropout or graduate. dropout is defined as leaving high school without completing the final grade (i.e., grade 12 in South Africa). Individuals who were still enrolled in high school at the end of the survey (i.e., the survivals) are right-censored. These individuals are however retained in the sample to avoid selection bias. Nonparametric hazard estimates presented in Figure 1 shows that the risk of dropping out of high school is higher than the risk of graduating during the first 13 years
of curriculum life. The risks of dropping out and graduating are the same in the 14th year. Figure 1 further shows that after the 14th year, the risk of graduating becomes higher than the risk of dropping out of high school.

Most developed countries in the world tend to have well structured and regulated curriculum that allow students to complete their schooling within the stipulated time. However, in a developing country setting like South Africa, the curriculum may not be well structured and strictly regulated. This means that students may spend more than the required time before completing their education. For example, more than 50% of graduates in my sample spent above the stipulated 12 years in education before obtaining a high school diploma (see Figure 2).

Figure 1: Nonparametric hazard estimates

This means that the stock sample of the first wave potentially over-samples individuals with too long duration of curriculum, leading to possible length bias in estimation. This requires some methodological adjustment to

\footnote{Given that curriculum is not strictly regulated in a developing country like South Africa, it is possible for a student to still drop out or graduate in the 13th year (or beyond) of curricula life. See Figure 2 for example.}

\footnote{The idea here is that anyone who has been able to survive in education up to the 14th year would rather want to push to graduate, instead of quitting at this stage.}
the standard competing risks model in order to account for possible bias. In section 3, I show how I make this adjustment. The distribution in Figure 2 shows only complete spells; censored spells are excluded.

The management of high school education in South Africa rests with the Department of Basic Education. The DBE is responsible for all schools from Grade R (i.e., grade 0 or the reception/preparatory year) to Grade 12 (the final year of high school). The South African Schools Act (SASA) of 1996 provides legal backing for access to school and the National Education Policy Act of 1996 regulates admission into public and independent (private) schools (Motala, Dieltiens, & Sayed, 2009).

Figure 2: curriculum distribution

Under apartheid, the Bantu system of Education (i.e., the official education system for black South Africans) severely limited the quality of education, and the apartheid regime consistently under-funded black schools. During the 1980s, there was considerable expansion of high schools for black learners, and by 1994 South Africa was able to provide broad access to basic education, although in a system characterized by racial inequality and offering poor quality to the majority of learners (Soudien, 2004). Whereas education
policy during apartheid provided for separate and different schooling for various racial groups, the focus of the post-apartheid democratic government has been to ensure equitable educational access for all based on the Constitution and Bill of Rights (Motala et al., 2009).

3 Method

3.1 Model structure

Every individual starts from high school enrollment and can transit to either of two possible destination states: high school dropout or high school graduation (see Figure 3). An individual who drops out of high school is no longer at “risk” of graduating and an individual who graduates is no longer at “risk” of dropping out of high school. Thus, high school drop out and high school graduation are competing risks.

![Figure 3: high school transition](image)

Survival and progression in high school education is measured annually. Dropouts are those who quit high school prior to completing the last grade (i.e., grade 12), while graduates are individuals who complete grade 12. Each year an enrolled student could either drop out or graduate. If none of these occurs, a student will continue in enrollment. Each student contributes only one spell; no repeated spells. That is, an individual who exits from school enrollment can only transit to a single destination (either dropout or graduation as in Figure 3).
The decision to select into dropout or graduation is determined by both observed and unobserved characteristics. That is,

\[ D_i = w_i \gamma + \epsilon_i \]  

\[ G_i = x_i \beta + \upsilon_i \]  

Where \( w_i \) and \( x_i \) are vectors of exogenous regressors, \( \gamma \) and \( \beta \) are vectors of parameters and \( \epsilon_i \) and \( \upsilon_i \) are unobserved components. I assume a discrete bivariate distribution of unobserved heterogeneity. For this distribution I suggest that destination-specific marginal distributions (i.e., \( \epsilon_i \) and \( \upsilon_i \)) have identical supports but potentially different probability mass values and unobserved heterogeneity has a mixing distribution with 2 points of support. I represent the points of support as either 0 or \( \mu \) across destinations (0 and \( \mu \) represents two different groups of individual based on unobserved factors. For example, 0 could represents low innate ability individuals, while \( \mu \) represents high innate ability individuals). Thus, there are four possible pairs that can be drawn from this bivariate distribution: (0, 0); (0, \( \mu \)); (\( \mu \), 0); (\( \mu \), \( \mu \)). Since both destination states are allowed to share the same support point in my model; let the first element in the pairs be dropout-specific draw, and the second element in the pairs be graduation-specific draw. Thus, the pattern of selection individuals exhibit is influenced by the combination of \( \mu \) and 0 that they drew, say from birth.

In order to access the pattern of self-selection into high school dropout and high school graduation (i.e., to test the hypothesis of selection on unobservables), I am particularly interested in the correlation between unobserved components in Equations (1) and (2). That is,

\[ \text{corr}(\upsilon_i, \epsilon_i) = r. \]  

I allow for general correlation between unobserved components.\(^4\) The

\(^4\)The typical assumption in many competing risks model is that unobserved components are only perfectly correlated across risks or destinations. This assumption may be too strong in practice. In my case, it suggests that the same unobserved component determines both dropout and graduation. This seems too restrictive.
correlation parameter, $r$ can either be positive, negative or zero. A statistically significant $r$ indicates the presence of self-selection. A positive and statistically significant $r$ suggests positive selection between dropping out and graduating. That is, the same sets of unobserved factors drive selection into high school dropout and high school graduation. This selection pattern means that an individual predominantly draw $(0,0)$ or $(\mu, \mu)$ from all possible pairs. A negative and statistically significant $r$ suggests a negative selection. That is, different sets of unobserved factors drive selection into dropout and graduation. This pattern of selection means that an individual predominantly draw $(0,\mu)$ or $(\mu,0)$ from all possible pairs. A statistically insignificant $r$ implies the absence of selection on unobservables.

Unlike in many developed countries, academic curriculum are not well structured and regulated in South Africa. As shown in Figure 2, more than 50% of high school graduates in my sample spent above the stipulated 12 years in education before obtaining a high school qualification. In order to avoid possible length bias due to oversampling of individuals with longer curriculum duration, I condition on the elapsed duration. That is, the competing risks of high school dropout and graduation are estimated, taking into account the duration an individual has already spent in high school education.

3.2 Econometric specification

To assess the pattern of self-selection into high school dropout and graduation, I set up a conditional discrete-time multivariate mixed proportional hazard (competing risks) model. Let $t_e$ (elapsed duration) denote the number of years an individual has spent in education prior to sampling, and $t_r$ (residual duration) the number of years spent in education after sampling. Let $T$ denote the entire duration spent in education and $t$ denote its realization, such that $t = t_e + t_r$. An individual can exit to a $\bar{y}$ mutually exclusive and exhaustive destination states.\footnote{Dropout or graduate in my case. A student who drops out is no longer at risk of graduating because the competing event (dropout) hinders the occurrence of the event of interest (graduation).} Let the random variable $Y, (Y = 1, \ldots, \bar{y})$
represent destination states. For ease of notation, \( t_e + t_r \) is replaced by \( t \). Thus, the exit process in terms of \( \bar{y} \) transition intensities at each point in time can be described as

\[
\theta_y(t) = P(T = t, Y = y|T \geq t),
\]  

(3)

We can think of a competing risks as a model in which the transition intensities are hazard functions of \( \bar{y} \) independent destination-specific latent duration or survival times (Dolton & van der Klaauw, 1999). More generally, the actual exit time and destination can be represented as

\[
T = \min(T_y : y = 1, ..., \bar{y})
\]

\[
Y = \text{argmin}_y(T_y : y = 1, ..., \bar{y})
\]

Each independent random variable \( T_y, y = 1, ..., \bar{y} \), is a latent duration that captures the length of stay before an exit of type \( y \) occurs in the absence of all other types of exit risks. The distribution \( f_y(t) \), conditional on observed and unobserved characteristics, can be parameterized with the proportional hazard

\[
\theta^y_i(t|x, v) = \lambda^y_i(t) \cdot \exp[x'_i \beta_y] \cdot v_y,
\]

(4)

where \( \lambda^y_i(t) \) is the risk-specific baseline hazard rate, \( x_i \) represents a set of explanatory variables (see Table 1), \( \beta_y \) is a vector of unknown parameters, while \( v_y \) is a destination-specific realization of the draw from the multivariate distribution of unobserved heterogeneity.

Let time be grouped into interval \( K^y_i = [k^y_{i0}, k^y_{i1}); [k^y_{i1}, k^y_{i2}) \ ... \ [k^y_{i-1}, \infty) \). The researcher is only able to observe time \( T \in \{1, ..., j\} \), where \( T = t \) denotes an exit within the interval \( [k^y_{i-1}, k^y_{i}) \). Therefore, the probability that a risk-specific transition has taken place within the interval \( [k^y_{i-1}, k^y_{i}) \), given that no transition occurred before \( k^y_{i-1} \) can be presented as

\[
h^y_i(t) = 1 - \exp \left( - \int_{k^y_{i-1}}^{k^y_{i}} \theta_y(s; \beta_y, v_y) ds_y \right),
\]

(5)
where \( s_y \) is used to capture the underlying continuous time. I can rewrite Equation (5) in a flexible complementary log-log form as

\[
h^y_i(t) = 1 - \exp \left( -\exp \left( \log \left( \int_{k_yt-1}^{k_yt} \lambda_y(s_y) \, ds_y \right) + x_i' \beta_y + \log(v_y) \right) \right) \tag{6}
\]

Unobserved heterogeneity and baseline hazard have unknown distributions. The flexibility of Equation (6) makes it possible for me to specify both the baseline hazard and unobserved heterogeneity non-parametrically. To specify the baseline hazard non-parametrically, I use a set of dummies \( \lambda_{yk} \), to characterize the continuous baseline as

\[
\lambda^y_k = \log \left( \int_{k_yt-1}^{k_yt} \lambda_y(s_y) \, ds_y \right)
\]

To further simplify notation, I define \( \log(v_y) = \mu_y \) and then rewrite Equation (6) as

\[
h^y_i(t) = 1 - \exp \left\{ -\exp \left\{ \lambda_{yk} + x_i' \beta_y + \mu_y \right\} \right\} \tag{7}
\]

Let \( g_i \) denote the censoring indicator for each destination state, where \( g_i = 1 \) if the spell is not censored and \( g_i = 0 \) otherwise. The overall destination-specific likelihood for student \( i \) who has a spell duration of \( k_yt \), conditional on the observed component \( x_i \) and the unobserved component \( v_y \) can be presented as

\[
l_i = (h^y_i(k_y; \beta_y; \mu_y))^{g_i} \cdot \prod_{s_y=1}^{k_yt-g_i} (1 - h^y_i(s_y; \beta_y; \mu_y)) \tag{8}
\]

I assume a discrete multivariate distribution of unobserved components. For this distribution I suggest that destination-specific marginal distributions have identical supports but potentially different probability mass values. Let there be \( Z \) mass points in the support, forming a set \( \Omega_\mu = \{ \mu^{[1]}, \mu^{[2]}, \ldots, \mu^{[Z]} \} \). Then for any individual a draw from the distribution of unobserved heterogeneity is a set of destination specific probability mass points \( \{ \mu_y \}_{y=1,\ldots,\bar{y}} \), where \( \mu_y \in \Omega_\mu \) for any \( y \), and \( \{ p_{yz} \}_{y=1,\ldots,\bar{y}, z=1,\ldots,Z} \) is the corresponding set of
probability mass values, $\sum_{y=1}^{\bar{y}} \sum_{z=1}^{Z} P_{yz} = 1$. Since $\mathbf{x}_i$ includes an intercept, normalization requires setting one element in $\Omega_\mu$ to zero. Conditioning on the elapsed duration in order to avoid length bias (see Figure 2), the total conditional log-likelihood to be maximized is presented as

$$L_i(k_y, \beta_y, \mu_y) = \sum_{i=1}^{N} \log \left[ \sum_{y=1}^{\bar{y}} \sum_{z=1}^{Z} P_{yz} \left( \left( h_i(k_y, \beta_y, \mu_y) \right)_{y \neq i}^{k_{yt-yi}} \prod_{s=1}^{k_{yt-sy}} \left( 1 - h_i(s_y; \beta_y, \mu_y) \right) \right) \right]$$

(9)

One of the main issues with competing risks model is identification. That is, to identify the joint distribution of latent failure times from the distribution of the identified minimum. Identification results and proof are provided by Heckman and Honoré (1989) and Abbring and Berg (2003). Generally, competing risks are identified by explanatory variables. Heckman and Honoré (1989) requires at least one explanatory variable with continuous variation from 0 to $\infty$ for identification. This is a stringent requirement because researchers do not often have explanatory variables that satisfy this condition. The results in Abbring and Berg (2003) relaxed this requirement. For identification, they require at least one covariate that has continuous variation within an interval. It is not the entire positive line as in Heckman and Honoré (1989), but just an open set. One of such variables in my case is reservation wage, which is sufficient for identification based on the results in Abbring and Berg (2003).

3.3 Analytical expression

Let $G$ and $D$ by two (discrete) random variables representing graduates and dropouts respectively. Thus, I am interested in the correlation that exists between $G$ and $D$. If I assume that all marginal distributions have identical mass points at the support, I can show (see Appendix A) that the expression
for the correlation coefficient is given as

\[
r = \frac{p_{11} - p_1^G p_1^D}{\sqrt{p_1^G(1-p_1^G)p_1^D(1-p_1^D)}}^{0.5}
\] (10)

where

\[
p_1^G = p_{10} + p_{11}
\]
\[
p_1^D = p_{01} + p_{11}
\]

The three parameters \((p_{01}, p_{10}, p_{11})\) required to compute the measure of selection are estimated in the competing risks model specified in Equation (9). In order to test whether the estimated measure of selection is statistically significant, I use the Delta Method to compute the variance of the correlation coefficient (see Appendix B for derivation).

4 Estimation results

Empirical results from the competing risks specification in Equation (9) are summarized in Table 2. Estimates in columns 1 and 2 do not account for the role of unobserved individual heterogeneity. The assumption here is that only observable characteristics are relevant in predicting high school dropout and graduation. The results in columns 3 and 4 account for a univariate distribution of unobserved heterogeneity, while estimates in columns 5 and 6 allow for a generalization of unobserved heterogeneity.

The results indicate that being a male increases the likelihood of dropping out of high school, \textit{ceteris paribus}. This makes sense in the context of a developing country like South Africa where male children are often called upon to take up menial jobs in order to supplement low family income. This result is consistent with the findings in Oreopoulos and Page (2006). Moving from the model without unobserved heterogeneity (column 2) to the model with a univariate distribution of unobserved differences (column 4) leads to substantial upward change in age effect, whereas generalizing unobserved heterogeneity (column 6) does change the age estimate significantly. How-
ever, I find no gender effect on graduation. That is, all things being equal, it will take the same amount of time for a man and a woman to graduate. Similarly, I find no location (urban or rural) or proximity-to-school effects across destination states.

As expected mother’s education and higher test scores reduce the likelihood of high school dropout. This is also consistent with previous studies in the literature (Eckstein & Wolpin, 1999; Li, Poirier, & Tobias, 2004; Foley et al., 2014). On the other hand, individuals with high test scores and whose mothers are more educated take longer time to graduate. This seemingly counter-intuitive result could be as a result of higher extracurricular activity related to education. This is especially so if higher-educated mothers are also correlated with wealthier families. Comparing columns 2, 4, and 6 for dropouts shows that the inclusion of a univariate unobserved heterogeneity does significantly change the results, while the generalization of unobserved factors leads to minor changes in the estimates. The same is true for graduates when I compare results in columns 1, 3 and 5.

Furthermore, teacher’s availability reduces the likelihood of high school dropout. As before, the introduction of a univariate distribution of unobserved heterogeneity leads to a substantial increase in the estimated coefficient, whereas a generalization of unobserved differences only cause a small change. However, I find no absenteeism effect on the duration until graduation. The time spent on household work and lack of help with homework both increase the likelihood of high school dropout and the duration until graduation for those who have graduated. In addition, while reservation
### Table 2: Competing risks model estimates

<table>
<thead>
<tr>
<th></th>
<th>Graduates</th>
<th>Dropout</th>
<th>Graduates</th>
<th>Dropout</th>
<th>Graduates</th>
<th>Dropout</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex (male)</td>
<td>0.110</td>
<td>0.284**</td>
<td>0.155</td>
<td>0.457**</td>
<td>0.182</td>
<td>0.460**</td>
</tr>
<tr>
<td></td>
<td>(0.076)</td>
<td>(0.067)</td>
<td>(0.086)</td>
<td>(0.097)</td>
<td>(0.105)</td>
<td>(0.096)</td>
</tr>
<tr>
<td>Location (urban)</td>
<td>-0.168</td>
<td>-0.244**</td>
<td>-0.174</td>
<td>-0.252</td>
<td>-0.179**</td>
<td>-0.233**</td>
</tr>
<tr>
<td></td>
<td>(0.107)</td>
<td>(0.084)</td>
<td>(0.115)</td>
<td>(0.154)</td>
<td>(0.029)</td>
<td>(0.110)</td>
</tr>
<tr>
<td>Mother’s education</td>
<td>-0.043**</td>
<td>-0.050**</td>
<td>-0.054**</td>
<td>-0.111**</td>
<td>-0.060**</td>
<td>-0.116**</td>
</tr>
<tr>
<td></td>
<td>(0.009)</td>
<td>(0.008)</td>
<td>(0.012)</td>
<td>(0.019)</td>
<td>(0.019)</td>
<td>(0.018)</td>
</tr>
<tr>
<td>Test scores</td>
<td>-0.030**</td>
<td>-0.081**</td>
<td>-0.042**</td>
<td>-0.113**</td>
<td>-0.045**</td>
<td>-0.110**</td>
</tr>
<tr>
<td></td>
<td>(0.008)</td>
<td>(0.007)</td>
<td>(0.010)</td>
<td>(0.011)</td>
<td>(0.013)</td>
<td>(0.010)</td>
</tr>
<tr>
<td>Distance to school</td>
<td>0.000</td>
<td>0.004</td>
<td>-0.000</td>
<td>0.007</td>
<td>-0.000</td>
<td>0.006</td>
</tr>
<tr>
<td></td>
<td>(0.002)</td>
<td>(0.002)</td>
<td>(0.002)</td>
<td>(0.003)</td>
<td>(0.000)</td>
<td>0.004</td>
</tr>
<tr>
<td>Teacher not absent</td>
<td>0.119</td>
<td>-0.250**</td>
<td>0.130</td>
<td>-0.377**</td>
<td>0.151</td>
<td>-0.388**</td>
</tr>
<tr>
<td></td>
<td>(0.120)</td>
<td>(0.087)</td>
<td>(0.143)</td>
<td>(0.133)</td>
<td>(0.086)</td>
<td>(0.124)</td>
</tr>
<tr>
<td>Time on household work</td>
<td>0.021**</td>
<td>0.025**</td>
<td>0.021**</td>
<td>0.035**</td>
<td>0.022</td>
<td>0.033**</td>
</tr>
<tr>
<td></td>
<td>(0.006)</td>
<td>(0.005)</td>
<td>(0.008)</td>
<td>(0.012)</td>
<td>(0.016)</td>
<td>(0.010)</td>
</tr>
<tr>
<td>Reservation wage</td>
<td>0.062</td>
<td>0.124**</td>
<td>0.063</td>
<td>0.126**</td>
<td>0.062</td>
<td>0.127**</td>
</tr>
<tr>
<td></td>
<td>(0.041)</td>
<td>(0.033)</td>
<td>(0.051)</td>
<td>(0.057)</td>
<td>(0.036)</td>
<td>(0.035)</td>
</tr>
<tr>
<td>No help with homework</td>
<td>0.300**</td>
<td>0.353**</td>
<td>0.336**</td>
<td>0.505**</td>
<td>0.345</td>
<td>0.493**</td>
</tr>
<tr>
<td></td>
<td>(0.100)</td>
<td>(0.094)</td>
<td>(0.111)</td>
<td>(0.116)</td>
<td>(0.193)</td>
<td>(0.125)</td>
</tr>
</tbody>
</table>

**Unobserved component**

- 1.448** (0.149)
- 1.477** (0.098)

**Probability Type 1**

- 0.1067** (0.038)
- 0.0381 (0.046)

**Probability Type 2**

- 0.893** (0.0385)
- 0.043 (0.043)

**Probability Type 3**

- 0.147** (0.073)

**Probability Type 4**

- 0.770** (0.066)

**Correlation coefficient ($r$)**

- 0.215 (0.155)

**Log-likelihood at maximum**

- -3414.87
- -3402.23
- -3401.09

**Sample size**

- 2,877
- 2,877
- 2,877

**Notes:** Standard errors are in parenthesis and ** represents statistical significance at 5%. UH: unobserved heterogeneity.
wage is positively associated with the likelihood of dropping out of high school, I find no effect on the likelihood of graduation. Overall, the inclusion of a univariate unobserved heterogeneity changes the estimated coefficients substantially, while generalization from a univariate distribution to a bivariate distribution of unobserved individual heterogeneity does not seem to add much.

In Figure 4, I plot the estimated flexible baseline hazards. The result indicates that the risk of dropping out of high school is higher than the risk of completing high school education in the first 12 years of schooling. The risks of dropping out and graduating are the same in the 13th year of education. Beyond the 13th year, graduating from high school becomes more likely than dropping out. Therefore, conditioning on observed and unobserved individual heterogeneity, the drop out and graduation pattern is consistent with the pattern observed in the data (see Figure 1).

In order to assess the patterns of self-selection, I turn to the competing risks estimates in columns 5 and 6 of Table 2. Particular interest is in the estimated correlation coefficient (a measure of selection on unobservables). Unlike the results in columns 3 and 4, the correlation coefficient incorporates
the bivariate distribution of unobserved heterogeneity between dropouts and graduates. This allows me to directly check for any systematic patterns of self-selection that may exist across destination states. The measure of selection is estimated as 0.215. Ignoring statistical significance for a moment, the estimated measure of self-selection suggests the presence of positive sorting. That is, individuals in our sample could either graduate faster or drop out faster based on unobserved factors. If an individual is good (e.g., high ability), they will be good in both dimensions, and if bad (e.g., low ability), they will be bad in both dimensions.

However, the estimated measure of self-selection is not statistically different from zero. This is important because it suggests that there is no systematic pattern of self-selection into high school dropout and graduation. That is, individuals who choose to drop out of high school are not systematically different from those who choose to complete high school education in terms of unobservables. It is typically found in developed country data that individuals who drop out of high school are those with relatively low ability, motivation, low expectations, and a set of negative preferences (Eckstein & Wolpin, 1999; Li et al., 2004; Foley et al., 2014). My results show that this is not necessarily the case in South Africa, a developing country in Sub-Saharan Africa. Individuals who drop out in my data do not seem to do so because they perceive or understand that they do not stand a chance in education. Rather, the factors that drive selection into dropout and graduation are observable individual and family characteristics. Thus, individuals who drop out of high school could potentially complete high school education like their counterparts with high school diploma.

What can possibly explain this finding? South Africa as a country still lives with the legacy of unequal opportunity after many years of apartheid. Inequality of opportunity implies that individuals with disadvantaged background may drop out of school more because the monetary and opportunity cost of completing high school education is higher for them. Also, South Africa, like many other developing countries, lacks viable social protection programmes or other forms of government support that could keep disadvantaged youths in school until graduation. Competing needs for the little
resources available to many poor households may not leave enough to support high school education. For example, the results in column 6 of Table 2 shows that low level of parental education (a variable that is positively correlated with family income) increases the likelihood of high school dropout. In addition, individuals with disadvantaged background are sometimes called upon to supplement low family income. This often implies dropping out of high school and taking up menial jobs.

This is not the kind of result one would expect to see in most developed countries where government support and social protection is relatively high. In such countries, some sort of negative sorting is most likely: individuals who drop out of high school are not potential high school graduates. They are different, either in their motivation, ability, expectations or preferences.

To the best of my knowledge, this paper presents the first evidence of self-selection into high school dropout and high school graduation in the literature. Previous attempts to account for unobserved differences in the decision to dropout of high school took a univariate approach. That is, researchers specify their model with a univariate distribution of unobserved heterogeneity. Given the restrictive nature of this approach, it does not allow for a quantification of self-selection in terms of unobservables. By allowing for a bivariate distribution of unobserved heterogeneity in my model, I am able to estimate a measure of selection.

In addition, the method I proposed in this paper provides an alternative, flexible and data-driven approach of assessing self-selection in education choices. The approach requires no sundry assumptions for identification. Rather, it simply uses the structure of a well-known model to estimate a measure of self-selection and document an empirical evidence.

5 Conclusion

This paper assesses the pattern of self-selection into high school dropout and graduation using longitudinal data from South Africa. Using the structure of competing risks model, I quantify the nature of self-selection and establish to what extent selection on unobservables matters. Results show that
there is no systematic pattern of self-selection into high school dropout and high school graduation. That is, individuals who choose to drop out of high school are not systematically different from those who choose to complete high school education in terms of unobservables. Rather, the factors that drive selection into dropout and graduation are observable individual and family characteristics. This suggests that individuals who drop out do not do so because they perceive or understand that they do not stand a chance in education. Thus, dropouts could potentially complete high school education.

Therefore, my results show that there is scope for government intervention in order to address the problem of high school dropout in South Africa, and possibly in other Sub-Saharan African countries. For example, disadvantaged students could be supported with social protection programmes like conditional cash transfers and other related incentive-based policies in order to continue in education. In addition, Individuals may not be aware of the benefit of a high school diploma in the labour market. Thus, providing accurate information may incentivize an individual to complete high school education.
References


Appendix

A Correlation coefficient derivation

Consider

$$Corr(G, D) = \frac{Cov(G, D)}{[Var(G)Var(D)]^{0.5}}$$

so that

$$Corr(G, D) = \frac{E(GD) - E(G)E(D)}{[Var(G)Var(D)]^{0.5}}$$

and

$$Corr(G, D) = \frac{E(GD) - E(G)E(D)}{[(E(G^2) - E(G)^2)(E(D^2) - E(D)^2)]^{0.5}}$$

Given the bivariate distribution of unobserved heterogeneity (2-points of support) below

$$\begin{bmatrix}
G, D & \mu_0 & \mu_1 \\
\mu_0 & p_{00} & p_{01} \\
\mu_1 & p_{10} & p_{11}
\end{bmatrix}$$

the marginal distribution for graduates can be represented as

$$\mu_0 : = p_{00} + p_{01}$$

$$\mu_1 : = p_{10} + p_{11}$$

and that of dropouts is given as

$$\mu_0 : = p_{00} + p_{10}$$

$$\mu_1 : = p_{01} + p_{11}$$

If I assume that all marginal distributions have identical mass points at the support, $Corr(G, D)$ can be presented in a more general form as follows:

$$Corr(G, D) = \frac{\sum_i \sum_j \mu_i \mu_j p_{ij} - \sum_i \mu_i p_i^G \sum_i \mu_i p_i^D}{[(\sum_i (\mu_i)^2 p_i^G)^2 - (\sum_i \mu_i p_i^G)^2)(\sum_i (\mu_i)^2 p_i^D - (\sum_i \mu_i p_i^D)^2)]^{0.5}}$$
Expanding, using the bivariate distribution of unobserved heterogeneity, I have that $\text{Corr}(G, D)$ is given as

$$
\left(\mu_0 \mu_0 p_{00} + \mu_0 \mu_1 p_{01} + \mu_1 \mu_0 p_{10} + \mu_1 \mu_1 p_{11}\right) - (\mu_0 p_G^0 + \mu_1 p_G^1)(\mu_0 p_D^0 + \mu_1 p_D^1)
\left[\left(\mu_0^2 p_G^0 + \mu_1^2 p_G^1\right) - (\mu_0 p_G^0 + \mu_1 p_G^1)^2\right]\left[\left(\mu_0^2 p_D^0 + \mu_1^2 p_D^1\right) - (\mu_0 p_D^0 + \mu_1 p_D^1)^2\right]^{0.5}
$$

Setting $\mu_0 = 0$ across destinations states (normalization assumption for identification purpose), I have that

$$
\text{Corr}(G, D) = \frac{(\mu_1 \mu_1 p_{11}) - (\mu_1 p_G^1)(\mu_1 p_D^1)}{\left[\left(\mu_1^2 p_G^1 - (\mu_1 p_G^1)^2\right)\left(\mu_1^2 p_D^1 - (\mu_1 p_D^1)^2\right)\right]^{0.5}}
$$

Rearranging and collecting terms, I get

$$
\text{Corr}(G, D) = \frac{\mu_1^2 (p_{11} - p_G^1 p_D^1)}{\left[\mu_1^2 \left(p_G^1 - (p_G^1)^2\right)\mu_1^2 \left(p_D^1 - (p_D^1)^2\right)\right]^{0.5}}
$$

And

$$
\text{Corr}(G, D) = \frac{\mu_1^2 (p_{11} - p_G^1 p_D^1)}{\mu_1^2 \left[\left(p_G^1 - (p_G^1)^2\right)\left(p_D^1 - (p_D^1)^2\right)\right]^{0.5}}
$$

Ultimately, the correlation coefficient between $G$ and $D$ writes as:

$$
r = \frac{p_{11} - p_G^1 p_D^1}{[p_G^1 (1 - p_G^1) p_D^1 (1 - p_D^1)]^{0.5}}
$$
B Delta method and variance estimation

The Delta Method (also known as the method of propagation of errors) provides an easily implementable method of computing the variance of a non-linear transformation of a random parameter vector. In my case, this is the estimated correlation coefficient $r$. Using the Delta Method, the variance of $r$ can be represented as:

$$V = JCJ^T,$$

where $J^T$ is the transpose of $J$ and $C$ is the covariance matrix of the estimated probability mass values (i.e., $p_{00}$, $p_{01}$, and $p_{10}$). $J$ is simply the matrix of first order derivatives of $r$, which is essentially a Jacobian matrix. That is:

$$J = \begin{bmatrix} \frac{\partial r}{\partial p_{00}} & \frac{\partial r}{\partial p_{01}} & \frac{\partial r}{\partial p_{10}} \end{bmatrix}.$$

The covariance matrix ($C$) comes directly from the estimated competing risks model. This can be represented as:

$$C = \begin{bmatrix} p_{001} & p_{01} & p_{10} \\ p_{002} & p_{012} & p_{102} \\ p_{003} & p_{013} & p_{103} \end{bmatrix}.$$

Thus, the variance of $r$ becomes

$$V = \frac{\partial r}{\partial p_{00}} \begin{bmatrix} p_{001} & p_{01} & p_{10} \\ p_{002} & p_{012} & p_{102} \\ p_{003} & p_{013} & p_{103} \end{bmatrix} \begin{bmatrix} \frac{\partial r}{\partial p_{00}} \\ \frac{\partial r}{\partial p_{01}} \\ \frac{\partial r}{\partial p_{10}} \end{bmatrix}.$$

To get $V$, I first need to solve for the partial derivatives of $r$ w.r.t the $p_s$. 

26
Recall the formula for correlation coefficient between dropouts and graduates is given as

\[
r = \frac{p_{11} - p_1^G p_1^D}{[p_1^G (1 - p_1^G) p_1^D (1 - p_1^D)]^{0.5}}
\]

where \( p_1^G = p_{10} + p_{11} \), and \( p_1^D = p_{01} + p_{11} \).

- Rearrangement of the expression for \( r \)

I estimate a parameter vector \([p_{00}, p_{01}, p_{10}]\), so for the purpose of computing the confidence interval of \( \hat{r} \) it is more convenient to consider the following substitution

\[
p_{11} = 1 - (p_{00} + p_{01} + p_{10})
\]

such that

\[
p_1^G = 1 - p_{00} - p_{01}
\]
\[
p_1^D = 1 - p_{00} - p_{10}
\]

and

\[
r = \frac{1 - (p_{00} + p_{01} + p_{10}) - p_1^G p_1^D}{[p_1^G (1 - p_1^G) p_1^D (1 - p_1^D)]^{0.5}}.
\]

This can be further rearranged as follows:

Consider

\[
p_1^G p_1^D = (1 - p_{00} - p_{01}) p_1^D = p_1^D - p_{00} p_1^D - p_{01} p_1^D
\]
\[
= [1 - p_{00} - p_{10}] - p_{00} p_1^D - p_{01} p_1^D.
\]
Inserting this into $r$ I have

$$
\begin{align*}
\frac{\partial}{\partial p_{00}} \left( \frac{p_{00}}{[p_1^G(1 - p_1^G)p_1^D(1 - p_1^D)]^{\frac{1}{2}}} \right) &= \\
= & \frac{[p_1^G(1 - p_1^G)p_1^D(1 - p_1^D)]^{\frac{1}{2}} - p_{00} \frac{1}{2} [p_1^G(1 - p_1^G)p_1^D(1 - p_1^D)]^{-\frac{1}{2}} \frac{\partial (p_1^G(1 - p_1^G)p_1^D(1 - p_1^D))}{\partial p_{00}}}{[p_1^G(1 - p_1^G)p_1^D(1 - p_1^D)]^{\frac{1}{2}}} \\
= & \frac{[p_1^G(1 - p_1^G)p_1^D(1 - p_1^D)]^{\frac{1}{2}} - p_{00} \frac{1}{2} [p_1^G(1 - p_1^G)p_1^D(1 - p_1^D)]^{-\frac{1}{2}} \frac{\partial (p_1^G(1 - p_1^G)p_1^D(1 - p_1^D))}{\partial p_{00}}}{[p_1^G(1 - p_1^G)p_1^D(1 - p_1^D)]^{\frac{1}{2}}} \\
= & \frac{1}{[p_1^G(1 - p_1^G)p_1^D(1 - p_1^D)]^{\frac{1}{2}}} - p_{00} \frac{1}{2} \frac{\partial (p_1^G(1 - p_1^G)p_1^D(1 - p_1^D))}{\partial p_{00}} \\
= & \frac{p_{00}}{[p_1^G(1 - p_1^G)p_1^D(1 - p_1^D)]^{\frac{1}{2}}} \left\{ \frac{p_1^G(1 - p_1^G)p_1^D(1 - p_1^D)}{p_{00}} - \frac{1}{2} \frac{\partial (p_1^G(1 - p_1^G)p_1^D(1 - p_1^D))}{\partial p_{00}} \right\}
\end{align*}
$$

B.1 Derivative of $r$

- Derivatives of $r$ w.r.t $p_{00}$

First

$$
\begin{align*}
\begin{align*}
r &= 1 - (p_{00} + p_{01} + p_{10}) - \{1 - p_{00} - p_{10}\} - p_{00}p_1^D - p_{01}p_1^D \\
&= -p_{01} + p_{00}p_1^D + p_{01}p_1^D \\
&= p_{00} \left[ 1 - 1 + p_1^D \right] - p_{01} \left[ 1 - p_1^D \right] \\
&= p_{00} - (1 - p_1^D) \{p_{00} + p_{01}\} \\
&= \frac{p_{00}}{[p_1^G(1 - p_1^G)p_1^D(1 - p_1^D)]^{\frac{1}{2}}} \left\{ (1 - p_1^D) \left( 1 - p_1^D \right) \right\}^{\frac{1}{2}}.
\end{align*}
\end{align*}
$$
and

\[
\frac{\partial}{\partial p_{00}} \left( \left[ \left( 1 - p_1^D \right) \left( 1 - p_1^G \right) \right]^{\frac{1}{2}} \right) = \frac{1}{2} \left( \frac{1 - p_1^D}{p_1^G p_1^D} \right)^{-\frac{1}{2}} \frac{\partial}{\partial p_{00}} \left( \frac{1 - p_1^D}{p_1^G p_1^D} \right)
\]

\[
= \frac{\left[ p_1^G p_1^D \right]^{\frac{1}{2}}}{\left[ 1 - p_1^D \right] \left[ 1 - p_1^G \right]} \frac{1}{2} \frac{\partial}{\partial p_{00}} \left( \frac{1 - p_1^D}{p_1^G p_1^D} \right)
\]

Now, since \( r = \frac{p_{11} - p_1^G p_1^D}{\left[ p_1^G (1 - p_1^G) p_1^D (1 - p_1^D) \right]^{0.5}} \) both partial derivatives can be written more compactly as

\[
\frac{\partial}{\partial p_{00}} \left( \frac{p_{00}^{p_1^G (1 - p_1^G) p_1^D (1 - p_1^D)}^{\frac{1}{2}}}{p_1^G (1 - p_1^G) p_1^D (1 - p_1^D)} \right)
\]

\[
= \frac{r^3 p_{00}^{p_1^G (1 - p_1^G) p_1^D (1 - p_1^D)}}{(p_{11} - p_1^G p_1^D)^3} \left\{ \left( \frac{p_{11} - p_1^G p_1^D}{r^2 p_{00}} \right)^2 - \frac{1}{2} \frac{\partial}{\partial p_{00}} \left( \frac{p_1^G (1 - p_1^G) p_1^D (1 - p_1^D)}{p_1^G p_1^D} \right) \right\}
\]

and

\[
\frac{\partial}{\partial p_{00}} \left( \left[ \left( 1 - p_1^D \right) \left( 1 - p_1^G \right) \right]^{\frac{1}{2}} \right) = \frac{r [p_1^G p_1^D]}{p_{11} - p_1^G p_1^D} \frac{1}{2} \frac{\partial}{\partial p_{00}} \left( \frac{1 - p_1^D}{p_1^G p_1^D} \right)
\]
so that

\[
\frac{\partial r}{\partial p_{00}} = \frac{r^3 p_{00}}{(p_{11} - p_1^D p_1^C)^2} \left\{ \frac{(p_{11} - p_1^C p_1^D)^2}{r^2 p_{00}} - \frac{1}{2} \frac{\partial \left( p_1^C (1 - p_1^C) p_1^D (1 - p_1^D) \right)}{\partial p_{00}} \right\}
\]

\[
= \frac{r [p_1^C p_1^D]}{p_{11} - p_1^C p_1^D} \left[ \frac{r^2 p_{00}}{2} \frac{\partial \left( p_1^C (1 - p_1^C) p_1^D (1 - p_1^D) \right)}{\partial p_{00}} \right]
\]

Consider first

\[
\frac{\partial \left( p_1^C (1 - p_1^C) p_1^D (1 - p_1^D) \right)}{\partial p_{00}}
\]

\[
= \frac{\partial \left( p_1^C - [p_1^C]^2 \right) (p_1^D - [p_1^D]^2)}{\partial p_{00}}
\]

\[
= (p_1^D - [p_1^D]^2) \frac{\partial}{\partial p_{00}} \left( p_1^C - [p_1^C]^2 \right) + (p_1^C - [p_1^C]^2) \frac{\partial}{\partial p_{00}} \left( p_1^D - [p_1^D]^2 \right)
\]

\[
= (p_1^D - [p_1^D]^2) (2p_1^C - 1) + (p_1^C - [p_1^C]^2) (2p_1^D - 1)
\]

\[
= p_1^D (1 - p_1^D) (2p_1^C - 1) + p_1^C (1 - p_1^C) (2p_1^D - 1)
\]

\[
= (1 - p_1^D)(1 - p_1^C) \left[ p_1^D \frac{2p_1^C - 1}{(1 - p_1^C)} + p_1^C \frac{2p_1^D - 1}{(1 - p_1^D)} \right]
\]

\[
= \frac{(p_{11} - p_1^C p_1^D)^2}{r^2} \left[ p_1^D \frac{2p_1^C - 1}{1 - p_1^C} + p_1^C \frac{2p_1^D - 1}{1 - p_1^D} \right]
\]
so that

\[
\frac{\partial r}{\partial p_{00}} = \frac{r}{p_{11} - p_1^G p_1^D} \left[ 1 - \frac{r^2 p_{00}}{2 (p_{11} - p_1^G p_1^D)^2} \frac{(p_{11} - p_1^G p_1^D)^2}{r^2} \left[ p_1^G \frac{2 p_1^G - 1}{1 - p_1^G} + p_1^G \frac{2 p_1^D - 1}{1 - p_1^D} \right] \right]
\]

\[
- \frac{p_1^C p_1^D}{2} \frac{\partial}{\partial p_{00}} \left( \frac{(1 - p_1^D) (1 - p_1^G)}{p_1^G p_1^D} \right)
\]

\[
= \frac{r}{p_{11} - p_1^G p_1^D} \left[ 1 - \frac{p_1^G}{2} \frac{2 p_1^G - 1}{1 - p_1^G} + \frac{2 p_1^D - 1}{1 - p_1^D} \right]
\]

Finally consider

\[
\frac{\partial}{\partial p_{00}} \left( \frac{(1 - p_1^D) (1 - p_1^G)}{p_1^C p_1^D} \right)
\]

\[
= \frac{p_1^C p_1^D \frac{\partial}{\partial p_{00}} \left( 1 - p_1^D + p_1^C p_1^D \right) - (1 - p_1^D) (1 - p_1^G) \frac{\partial}{\partial p_{00}} (p_1^C p_1^D)}{[p_1^C p_1^D]^2}
\]

\[
= \frac{p_1^C p_1^D \left( 2 + \frac{\partial}{\partial p_{00}} (p_1^C p_1^D) \right) - (1 - p_1^D) (1 - p_1^G) \frac{\partial}{\partial p_{00}} (p_1^C p_1^D)}{[p_1^C p_1^D]^2}
\]

\[
= \frac{2 p_1^C p_1^D + \left[ p_1^C p_1^D - (1 - p_1^D + p_1^C p_1^D) \right] \frac{\partial}{\partial p_{00}} (p_1^C p_1^D)}{[p_1^C p_1^D]^2}
\]

\[
= \frac{2 p_1^C p_1^D + [p_1^D + p_1^G - 1] \frac{\partial}{\partial p_{00}} (p_1^C p_1^D)}{[p_1^C p_1^D]^2}
\]

\[
= \frac{2 p_1^C p_1^D + [p_1^D + p_1^G - 1] \{- p_1^D - p_1^G\}}{[p_1^C p_1^D]^2}
\]

\[
= \frac{2 p_1^C p_1^D - \left[ (p_1^D + p_1^G)^2 - (p_1^D + p_1^G) \right]}{[p_1^C p_1^D]^2}
\]

\[
= \frac{p_1^D - [p_1^D]^2 - p_1^G - [p_1^G]^2}{[p_1^C p_1^D]^2}
\]

\[
= \frac{p_1^D (1 - p_1^D) + p_1^G (1 - p_1^G)}{[p_1^C p_1^D]^2}
\]
such that

\[
\frac{\partial r}{\partial p_{00}} = \frac{r}{p_{11} - p_{11}^G p_{11}^D} \left[ 1 - \frac{p_{00}}{2} \left[ p_{11}^D 2p_{11}^G - 1 + p_{11}^G 2p_{11}^D - 1 \right] - \frac{p_{11}^D p_{11}^G (1 - p_{11}^G) + p_{11}^G (1 - p_{11}^D)}{2 [p_{11}^G p_{11}^D]^{\frac{3}{2}}} \right]
\]

\[
= \frac{r}{p_{11} - p_{11}^G p_{11}^D} \left[ 1 - \frac{p_{00}}{2} \left[ p_{11}^D 2p_{11}^G - 1 + p_{11}^G 2p_{11}^D - 1 \right] - \frac{1}{2} \frac{p_{11}^D (1 - p_{11}^D) + p_{11}^G (1 - p_{11}^G)}{p_{11}^G p_{11}^D} \right]
\]

\[
= \frac{r}{p_{11} - p_{11}^G p_{11}^D} \left[ 1 - \frac{p_{00}}{2} \left[ p_{11}^D \left( \frac{p_{11}^G - (1 - p_{11}^G)}{1 - p_{11}^G} \right) + p_{11}^G \left( \frac{p_{11}^D - (1 - p_{11}^D)}{1 - p_{11}^G} \right) \right] - \frac{1}{2} \left[ \frac{1 - p_{11}^D}{p_{11}^G} + \frac{1 - p_{11}^G}{p_{11}^D} \right] \right]
\]

- Derivatives of \( r \) w.r.t \( p_{01} \)

First

\[
\frac{\partial}{\partial p_{01}} \left( \frac{p_{00}}{[p_{11}^G (1 - p_{11}^G)p_{11}^D (1 - p_{11}^D)]^{\frac{3}{2}}} \right) = - \frac{p_{00} p_{11}^D (1 - p_{11}^D)}{2 [p_{11}^G (1 - p_{11}^G)p_{11}^D (1 - p_{11}^D)]^{\frac{3}{2}}} \frac{\partial}{\partial p_{01}} \left( p_{11}^G - [p_{11}^G]^{\frac{3}{2}} \right)
\]

\[
= - \frac{p_{00} p_{11}^D (1 - p_{11}^D)}{2 [p_{11}^G (1 - p_{11}^G)p_{11}^D (1 - p_{11}^D)]^{\frac{3}{2}}} \left( 2p_{11}^G - 1 \right)
\]

\[
= \frac{p_{00}}{[p_{11}^G (1 - p_{11}^G)p_{11}^D (1 - p_{11}^D)]^{\frac{3}{2}}} \frac{p_{11}^D (1 - p_{11}^D)}{2 [p_{11}^G (1 - p_{11}^G)p_{11}^D (1 - p_{11}^D)]^{\frac{3}{2}}} \left( 2p_{11}^G - 1 \right)
\]

\[
= \left( r + \left[ \frac{(1 - p_{11}^D)(1 - p_{11}^G)}{p_{11}^G p_{11}^D} \right]^{\frac{3}{2}} \right) \frac{2p_{11}^G - 1}{2p_{11}^G (1 - p_{11}^G)}
\]

32
and

$$\frac{\partial}{\partial p_{01}} \left( \left[ \frac{(1 - p^D_1) (1 - p^G_1)}{p^G_1 p^D_1} \right]^{\frac{1}{2}} \right)$$

$$= \frac{1}{2} \left[ \frac{(1 - p^D_1) (1 - p^G_1)}{p^G_1 p^D_1} \right]^{\frac{1}{2}} \frac{1 - p^D_1}{p^D_1} \frac{\partial}{\partial p_{01}} \left( \frac{1 - p^G_1}{p^G_1} \right)$$

$$= \frac{1}{2} \left[ \frac{(1 - p^D_1) (1 - p^G_1)}{p^G_1 p^D_1} \right]^{\frac{1}{2}} \frac{1 - p^D_1}{p^D_1} \left( -\frac{1}{[p^G_1]^2} \right)$$

$$= -\frac{1}{2} \frac{1}{p^D_1} \frac{1 - p^D_1}{p^G_1}$$

Thus,

$$\frac{\partial r}{\partial p_{01}} = \left( r + \left[ \frac{(1 - p^D_1) (1 - p^G_1)}{p^G_1 p^D_1} \right]^{\frac{1}{2}} \right) \frac{2p^G_1 - 1}{2p^G_1(1 - p^G_1)} + 1 \frac{r}{2} \frac{1 - p^D_1}{p^D_1 - p^G_1 p^D_1}$$

$$= \frac{r}{2p^G_1} \left[ \left( \frac{1 - p^G_1 (1 - p^D_1)}{p^D_1 - p^G_1 p^D_1} + 1 \right) \left[ \frac{p^G_1}{1 - p^G_1} - 1 \right] + \frac{1 - p^D_1}{p^D_1 - p^G_1 p^D_1} \right]$$

$$= \frac{r}{2p^G_1} \left[ \left( \frac{1 - p^G_1 (1 - p^D_1)}{p^D_1 - p^G_1 p^D_1} + 1 \right) \left[ \frac{p^G_1}{1 - p^G_1} - 1 \right] + \frac{1 - p^D_1}{p^D_1 - p^G_1 p^D_1} \right]$$

$$= \frac{r}{2p^G_1} \left[ \left( \frac{1 - p^G_1 - p^D_1 + p_{11}}{p^D_1 - p^G_1 p^D_1} \right) \left[ \frac{p^G_1}{1 - p^G_1} - 1 \right] + \frac{1 - p^D_1}{p^D_1 - p^G_1 p^D_1} \right]$$

$$= \frac{r}{2p^G_1} \left[ \left( \frac{1 - p^G_1 - p^D_1 + p_{11}}{p^D_1 - p^G_1 p^D_1} \right) \left[ \frac{p^G_1}{1 - p^G_1} - 1 \right] + \frac{1 - p^D_1}{p^D_1 - p^G_1 p^D_1} \right]$$

$$\frac{\partial r}{\partial p_{01}} = \frac{r}{2p^G_1 (p^D_1 - p^G_1 p^D_1)} \left[ p^G_1 \left[ \frac{p^G_1}{1 - p^G_1} - 1 \right] + (1 - p^D_1) \right].$$

Derivative of $r$ with respect to $p_{10}$ is analogous to that of $p_{01}$. The only difference is the indexing. With these partial deviates, I compute the variance.
of the correlation coefficient using the values of $p_{00}, p_{01}, p_{10}, p_{11}$ obtained from the competing risks model specified in Equation (9).