Assignment 2: Microeconometrics

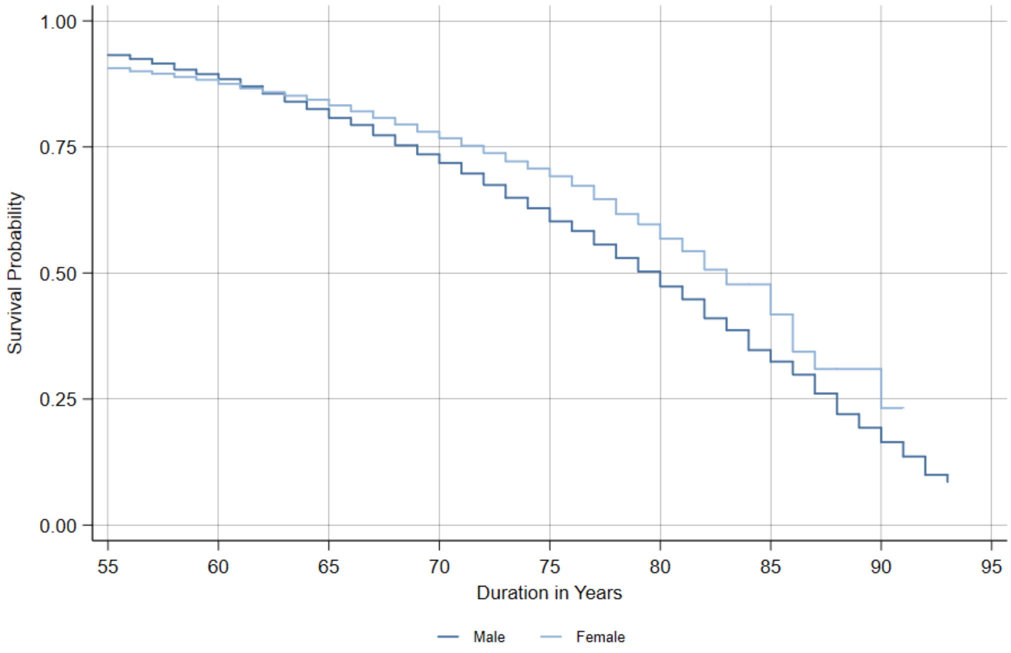
Daniel Redel (210630)
Instructor: Dr. Alexandros Theloudis

Autumn 2022

1 Survival Analysis

d. Figure 1 shows the Kaplan-Meier Survivor Estimator by gender and starting at age 55:

Figure 1: Kaplan-Meier Survivor Estimator, by Gender



The horizontal axis displays the age and the vertical axis the cumulative survival probability, $\hat{S}_T(t)$, or the proportion of people that are still alive after a stated age. Intuitively, this figure can be interpreted as the probability of surviving up until a given age. As expected, the function starts at one and monotonically declines to zero, indicating that all the individuals in the sample will eventually die. This is a decreasing step function with a jump at each discrete failure time. Note also that the survival curve decreases at **increasing rates** —over time, people become more likely to die—, reflecting a positive duration dependence on the spell. Finally, comparing the survival probability between males and females we can see that, from about the age of 65, females have a slightly higher probability of survival at a given age.

e. Table 1 displays the results of different Proportional Hazard specifications, specifying a Weibull distribution:

Table 1: Survival Regression Results - Hazard Rates

	(1)	(2)	(3)	(4)	(5)
_t					
Female		0.644***	0.638***	0.666***	0.665***
		(0.022)	(0.022)	(0.024)	(0.025)
Black		1.316***	1.279***	1.240***	1.218***
		(0.055)	(0.055)	(0.054)	(0.054)
Hispanic		0.751***	0.747***	0.789***	0.867**
		(0.047)	(0.048)	(0.051)	(0.056)
High School		0.693***	0.701***	0.790***	0.841***
		(0.026)	(0.027)	(0.031)	(0.033)
College		0.471***	0.480***	0.639***	0.714***
		(0.025)	(0.026)	(0.035)	(0.040)
Father			1.060	1.034	1.012
			(0.039)	(0.038)	(0.037)
Mother			1.096**	1.082**	1.023
			(0.041)	(0.040)	(0.038)
Current Smoker				2.853***	2.658***
				(0.129)	(0.122)
Past Smoker				1.453***	1.328***
				(0.067)	(0.061)
Heavy Drinker				1.002	1.130*
				(0.066)	(0.075)
Moderate Drinker				0.710***	0.804***
				(0.026)	(0.029)
Overweight				0.969	0.898***
				(0.039)	(0.036)
Obese				1.310***	1.046
				(0.059)	(0.049)
High Blood Pressure					1.397***
					(0.051)
Diabetes					2.245***
					(0.101)
Cancer					1.985***
					(0.130)
Lung Disease					2.044***
					(0.113)
Heart Problems					1.634***
					(0.073)
Arthritis					1.065*
					(0.038)
α	7.060***	6.629***	6.651***	7.130***	6.993***
N	12,613	12,602	12,262	12,262	12,262

Exponentiated coefficients

This table reports the results in terms of hazard metrics, $\exp(\beta_i)$, of each independent variable (to see the same results in terms of β coefficients, see Appendix). We can make the following observations:

- **Baseline Hazard - Column (1)**: The basic model without covariates reports the fitted α and the baseline hazard rates $\lambda_0(t, \alpha)$ (See the intercept term or constant in Appendix). The fit of the Weibull model exhibits positive state dependence since $\alpha = 7.06 > 1$. Thus, the probability of the spell terminating increases as the spell lengthens. All the models reached similar conclusions and are in line with what the shape of the Kaplan-Meier estimator suggested.

- **Basic Covariates - Column (2)**: All the individual characteristics added here resulted in statistically significant coefficients at 1% level, but only the variable associated with race increases the baseline hazard. More concretely, being black increases the hazard by 31.6%, relative to non-black individuals, which means a higher mortality rate. On the other hand, being female is associated with a reduction in the hazard by a factor of around 0.64, relative to males. In that sense, being female is associated with good prognostic.

- **Parents Longevity - Column (3)**: Adding information about their parents does not change the conclusions regarding our previous covariates in (1). On the other hand, while we see higher hazard rates associated with parents that had lower life expectancy, these results are only statistically significant in relation to the mother's longevity, not the father's. Individuals that had a mother that died before the age of 70 are associated with an increase in the hazard rate of about 10%.

- **Health Behavior - Column (4)**: Here we see different results depending on the type of health behavior. First, current smokers —and also past smokers to a less degree— have a strong and statistically significant correlation with higher hazard rates. Being obese increases the baseline mortality rate by 31% relative to not being obese (nor overweight). Interestingly, being a heavy drinker and having overweight does not have statistically significant results in the hazard rates. All the conclusions from the previous covariates in (2) and (3) remain.

- **Chronic Health - Column (5)**: Finally, all the chronic health conditions added in Column (4) increase the hazard rates, all statistically significant at the 1% level. Note that now having a low-longevity mother does not report significant results, which could mean that genetic characteristics were previously correlated to mothers' longevity. This could also be the case regarding the correlation between being obese —where now we see no significant "effects"— and these chronic conditions. The conclusions of the remaining covariates are similar to the conclusions from (2) to (4).

f. In a proportional hazard model (PH), the conditional hazard rate $\lambda(t|\mathbf{x})$ can be written as the product of two different functions: (i) $\lambda_0(t, \alpha)$, which is a baseline hazard function that does no depend on the individual characteristics $\mathbf{x}_i$; and (ii) $\phi(\mathbf{x}, \beta)$, an individual-specific function that describes the effects of $\mathbf{x}_i$:

$$\lambda(t|\mathbf{x}) = \lambda_0(t, \alpha)\phi(\mathbf{x}, \beta)$$

The `streg` command for PH modles specifies $\lambda_0(t, \alpha)$ to be parametric. In our case, we assume the hazard function to follow a Weibull distribution, defined as:

$$\lambda_T(t|\mathbf{x}) = \gamma\alpha t^{\alpha-1} \text{ and } S_T(t) = \exp(-\gamma t^{\alpha})$$

With $\gamma = \exp(\mathbf{x}'\beta)$, so that we end up with the following **hazard function**:

$$\lambda_T(t|\mathbf{x}) = \gamma\alpha t^{\alpha-1} = \exp(x'\beta)\alpha t^{\alpha-1}$$

To estimate this we need to rely on Maximum Likelihood Estimation. Additionally, we are in the presence of right-censored data, and MLE needs to adjust it's likelihood function to take that into account. In particular, the contribution to the likelihood function of individual i is given by: (i) the conditional density of t_i if the observation is not censored, or (ii) the conditional probability that $t_i > c_i$ in the case of censoring, so that the likelihood function is:

$$f(t|\mathbf{x}, \theta)^{\delta_i} S(t|\mathbf{x}, \theta)^{1-\delta_i}$$

where $\delta_i = 1$ when there is no censoring and $\delta_i = 0$ when right-censoring. For the uncensored data, the contribution to the likelihood in the context of a Weibull distribution is:

$$\begin{aligned} \ln f(t_i|\mathbf{x}_i, \beta, \alpha) &= \ln[\exp(\mathbf{x}'\beta)\alpha t^{\alpha-1} \exp(-\exp(\mathbf{x}'\beta)t^{\alpha})] \\ &= \mathbf{x}'\beta + \ln\alpha + (\alpha - 1)\ln t - \exp(\mathbf{x}'\beta)t^{\alpha} \end{aligned}$$

For the censored data, the likelihood contribution is the survival function:

$$\begin{aligned} \ln S(t_i|\mathbf{x}_i, \beta, \alpha) &= \ln[\exp(-\exp(\mathbf{x}'\beta)t^{\alpha})] \\ &= -\exp(x'\beta)t^{\alpha} \end{aligned}$$

Taking the logs, we have that the MLE $\hat{\theta}$ maximizes the following log-likelihood:

$$\ln L(\theta) = \sum_{i=1}^{N}[\delta_i \ln f(t_i|\mathbf{x}_i, \theta_i) + (1 - \delta_i)\ln S(t_i|\mathbf{x}_i, \theta_i)]$$

Which, with a little algebra, the log-likelihood function becomes:

$$\ln L(\theta) = \sum_{i=1}^{N}[\delta_i[\mathbf{x}'\beta + \ln\alpha + (\alpha - 1)\ln t - \exp(\mathbf{x}'\beta)t^{\alpha}] - (1 - \delta_i)\exp(x'\beta)t^{\alpha}]$$

g. As already stated, the PH model can be written as the product of:

$$\lambda(t|\mathbf{x}) = \lambda_0(t,\alpha)\phi(\mathbf{x},$$

Regarding the term on the left —the baseline hazard—, the first assumption about this model is that the hazard function follows a Weibull distribution, an assumption that is typically suitable for modeling data with monotone hazard rates that either increase or decrease exponentially with time. More concretely, the advantage of Weibull is that it allows for duration dependence, where:

1. If $\alpha > 1$, the hazard rate is monotonically **increasing**.
2. If $\alpha < 1$, the hazard rate is monotonically **decreasing**.

From Table 1 we can see that in each specification α is greater than 1, which means that the hazard of failure increases with time, so the shape that Weibull's distribution provides here is justified.

With respect to the proportionality conditions, these models assume that the shape of the survivor curve —which models the duration dependence— is common across individuals, but the individual characteristics defined in $\phi(\mathbf{x},\beta)$ affects the resulting λ proportionally. This means that all hazard functions $\lambda(t|\mathbf{x})$ are proportional to the baseline hazard $\lambda_0(t,\alpha)$, with some scale factor $\phi(\mathbf{x},\beta)$. In our case, $\exp(x_i'\beta)$ is the adjustment factor that depends on $\mathbf{x}_i$, but this adjustment is the same at all durations t. In fact, changes in regressors have the effect of a multiplicative change in the hazard function:

$$\frac{\partial\lambda(t)}{\partial\mathbf{x}} = \exp(\mathbf{x}'\beta)\alpha t^{\alpha-1}\beta = \lambda(t)\beta$$

Finally, regarding heterogeneity, PH models can account for observed heterogeneity between individuals, in the sense we have just stated: individual characteristics $\mathbf{x}$ have an impact that is proportional on the baseline hazard. However, this models do not take into account unobserved heterogeneity v_i, where individuals can differ in a way not fully accounted for by the observed covariates.

2 Generalized Method of Moments

Let's consider the following linear regression:

$$y_i = \mathbf{x}_i'\beta + u_i$$

where $\mathbf{x}_i' = [1, x_{1i}, x_{2i}, ..., x_{qi}]_{1\times q}$ the vector of q regressors and $\beta = [\beta_0, \beta_1, ..., \beta_q]_{q\times 1}$ their corresponding coefficients to estimate.

a. In OLS models, the **Strict Exogeneity** assumption, $E[\varepsilon_i|\mathbf{x}_i] = 0$, also implies **Orthogonality** between the error terms and the independent variables[1], $E[\mathbf{x}_i\varepsilon_i] = 0$. With this, we can derive $r = q$ population moment conditions:

$$E[\mathbf{x}_i(y_i - \mathbf{x}_i'\beta)] = 0$$

We then replace the population expectation with their corresponding sample analog, which results in the following q sample moment conditions:

$$\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i(y_i - \mathbf{x}_i'\beta) = 0$$

b. The GMM Estimator $\hat{\theta}_{GMM}$ minimizes the following objective function:

$$\hat{\theta} = \arg\min_{\theta} Q(\theta) = \left[\frac{1}{N}\sum_{i=1}^{N}\mathbf{h}(\mathbf{w}_i, \hat{\theta})\right]' W_N \left[\frac{1}{N}\sum_{i=1}^{N}\mathbf{h}(\mathbf{w}_i, \hat{\theta})\right]$$

where $\mathbf{h}(\cdot)$ is a $r \times 1$ vector that describes some relationship between the parameters θ and our data $\mathbf{w}_i = (\mathbf{y}_i', \mathbf{x}_i', \mathbf{z}_i')'$. In our case, we are exploiting the sample moment conditions to define this relationship. The GMM estimator is the solution of a minimization problem, which means that the parameter $\hat{\theta}$ is chosen so that the quadratic form of the analogic sample moment condition is as close to zero as possible. W_N is a $q \times q$ positive definite weighting matrix. We will discuss the choice of weighting matrix in the next question regarding optimal GMM.

$$Q(\theta) = \left[\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i(y_i - \mathbf{x}_i'\hat{\beta})\right]' W_N \left[\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i(y_i - \mathbf{x}_i'\hat{\beta})\right]$$

c. When the model is said to be **just-identified**, that is, when the number of moment conditions equals the number of unknown parameters to estimate $r = q$, the $\hat{\beta}_{GMM}$ has a closed-form solution and the Method of Moments (MM) estimator can be used:

$$Q(\theta) = \left[\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i(y_i - \mathbf{x}_i'\hat{\beta})\right]' \left[\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i(y_i - \mathbf{x}_i'\hat{\beta})\right]$$

[1] Recall that, by the law of iterated expectations (LIE): $E[\mathbf{x}_i\varepsilon_i] = E_{\mathbf{x}}[E(\mathbf{x}_i\varepsilon_i|\mathbf{x}_i)] = E_{\mathbf{x}}[\mathbf{x}_iE(\varepsilon_i|\mathbf{x}_i)] = E_{\mathbf{x}}[\mathbf{x}_i0] = 0$

In the exact-identification case, the weighting matrix W_N is redundant because the objective function $Q(\hat{\theta}_{MM})$ can converge to exactly zero. We can just define $W_N = \mathbf{I}_q$ Finally, the parameter $\hat{\beta}$ that solves this minimizing problem is:

$$\hat{\beta}_{MM} = \left(\sum_{i=1}^{N} \mathbf{x}_i \mathbf{x}_i'\right)^{-1} \sum_{i=1}^{N} \mathbf{x}_i y_i$$

This MM estimator is equivalent to the traditional OLS estimator. Thus, we can interpret OLS parameters as not only minimizing the sum of squared error but also achieving orthogonality between the error terms and data.

d. Let's now assume that regressors $\mathbf{x}_i$ are endogenous, so now we can't use the moment conditions previously derived. Alternatively, we have a vector of r relevant instruments $\mathbf{z}_i$ that satisfies $E[\varepsilon_i|\mathbf{z}_i] = 0$. The population moment conditions are now defined by:

$$E[\mathbf{z}_i\varepsilon_i] = E[\mathbf{z}_i(y_i - \mathbf{x}_i'\beta)] = 0$$

Replacing the population expectation with their corresponding sample analog results in the following r sample moment conditions:

$$\frac{1}{N}\sum_{i=1}^{N} \mathbf{z}_i(y_i - \mathbf{x}_i'\beta) = 0$$

Consequently, GMM will minimize the following objective function:

$$Q(\theta) = \left[\frac{1}{N}\sum_{i=1}^{N} \mathbf{z}_i(y_i - \mathbf{x}_i'\hat{\beta})\right]' W_N \left[\frac{1}{N}\sum_{i=1}^{N} \mathbf{z}_i(y_i - \mathbf{x}_i'\hat{\beta})\right]$$

Because the r instruments are functions of the q parameters —having at least as many instruments as independent variables $r \geq q$—, there could be a problem of **over-identification**. Note also that this objective function depends on the choice of weighting matrix, W_N, so the GMM estimator also depends on the weighting matrix. Finally, the parameter $\hat{\beta}$ that solves this minimizing problem is:

$$\hat{\beta}_{GMM} = [X'ZW_NZ'X]^{-1}X'ZW_NZ'Y$$

As we can see, the use of a different weighting matrix will results in a different GMM estimator.

e. When $r > q$ the model is said to be **over-identified** and $Q(\theta)$ has no closed-form solution for θ as there are more moment equations (r) than unknowns (q). Instead, $\hat{\theta}$ is chosen so that a quadratic form of the analogic sample moment condition is as close to zero as possible.

Regarding the conditions for identification, θ is identified by the Rank Condition assumption:

Rank Condition: $G_0 = \operatorname{plim} G_N(\theta_0)$ of is a full-rank matrix

In the IV case, the rank condition means that plim $N^{-1}Z'X$ (or that the $r \times q$ matrix $E[\mathbf{x}_i\mathbf{z}_i']$) is of rank q. This ensures that the inverse of $Z'X$ exists, provided that W_N is of full rank. In that case, the weighting matrix W_N is a $r \times r$ matrix.

A necessary but not sufficient condition for identification is the **Order Condition** that $r \geq q$:

1. If $r = q$ the model is **just identified**.
2. If $r > q$ the model is **over identified**.

In the just-identified case $r = q$, for example, we see that $Z'Z$ is a square matrix that is invertible. Then:

$$[X'ZW_NZ'X]^{-1} = (Z'X)^{-1}W_N^{-1}(Z'X)^{-1}$$

and our GMM estimator $\hat{\beta}_{GMM}$ defined above simplifies to the **instrumental variables** estimator:

$$\hat{\beta}_{GMM} = [X'Z]^{-1}Z'Y$$

For overidentified models, however, the most efficient GMM estimator is the one that chooses the optimal choice of weighting matrix of $W_N = \hat{S}^{-1}$. This idea is discussed in the next question.

f. The weighting matrix W_N tells us how much weight to attach to which linear combinations of the sample moment. Different weighting matrices W_N lead to different consistent estimators with generally different asymptotic covariance matrices, which means that the choice of W_N is important in terms of efficiency. Using a general W_N for this case does not yield necessarily an efficient estimator:

$$V(\hat{\beta}_{GMM}) = N^{-1}(X'ZW_NZ'X)^{-1}(X'ZW_N\hat{S}W_NZ'X)(X'ZW_NZ'X)^{-1}$$

So, we would like to choose a W_N one that has the smallest asymptotic variance. This is often called optimal or efficient GMM. For our IV case with over-identification $r > q$, the **optimal GMM estimator or two-step GMM estimator** is defined by:

$$\hat{\beta}_{OGMM} = (X'Z\hat{S}^{-1}Z'X)^{-1}X'Z\hat{S}^{-1}Z'Y$$

with $W_N = \hat{S}^{-1}$ known as the most efficient GMM estimator for these cases. Then, their corresponding variance is:

$$V(\hat{\beta}_{OGMM}) = N^{-1}(X'Z\hat{S}^{-1}Z'X)^{-1}$$

Both the optimal GMM and the 2SLS estimator lead to efficiency gains in overidentified models. Optimal GMM has the advantage of being more efficient than 2SLS when in presence of heteroskedasticity. In those cases, $\hat{S}$ is:

$$\hat{S} = \frac{1}{N}\sum_{i=1}^{N}\hat{u}_i^2\mathbf{z}_i\mathbf{z}_i$$

where $\hat{u}_i$ are the GMM residuals.

Appendix

Table 2: Survival Regression Results - Coefficients

	(1)	(2)	(3)	(4)	(5)
_t					
Female		-0.439***	-0.449***	-0.407***	-0.408***
		(0.034)	(0.035)	(0.036)	(0.037)
Black		0.275***	0.246***	0.215***	0.197***
		(0.042)	(0.043)	(0.043)	(0.044)
Hispanic		-0.286***	-0.292***	-0.237***	-0.143**
		(0.062)	(0.064)	(0.064)	(0.065)
High school		-0.367***	-0.356***	-0.236***	-0.173***
		(0.038)	(0.039)	(0.039)	(0.039)
College		-0.753***	-0.734***	-0.447***	-0.336***
		(0.053)	(0.054)	(0.055)	(0.056)
Father			0.058	0.033	0.012
			(0.037)	(0.037)	(0.037)
Mother			0.092**	0.079**	0.023
			(0.037)	(0.037)	(0.037)
Current Smoker				1.048***	0.978***
				(0.045)	(0.046)
Past Smoker				0.374***	0.284***
				(0.046)	(0.046)
Heavy Drinker				0.002	0.122*
				(0.066)	(0.067)
Moderate Drinker				-0.342***	-0.218***
				(0.036)	(0.037)
Overweight				-0.032	-0.108***
				(0.040)	(0.040)
Obese				0.270***	0.045
				(0.045)	(0.047)
High Blood Pressure					0.334***
					(0.036)
Diabetes					0.809***
					(0.045)
Cancer					0.686***
					(0.065)
Lung Disease					0.715***
					(0.055)
Heart Problems					0.491***
					(0.044)
Arthritis					0.063*
					(0.036)
Const.	-31.406***	-29.029***	-29.173***	-31.732***	-31.510***
	(0.688)	(0.700)	(0.714)	(0.728)	(0.729)
α	7.060	6.629	6.651	7.130	6.993
N	12,613	12,602	12,262	12,262	12,262