Survival Analysis of Transfer Students
Franklin Zhou, Margot Neverett, Beverly King, Yihui Li, Kyle Chapman
East Carolina University

NCAIR 2023 - 50th Annual Conference
Back Together Again
LEARNING OBJECTIVES

• **Describe the importance to higher education institutions of investigating the factors that contribute to transfer student success;**

• **Identify factors that past research and the current study have determined influence the likelihood of drop out for transfer students;** and

• **Recognize how the survival analysis statistical technique can be used to assist in the understanding of transfer student success.**
BACKGROUND OF THE STUDY

- Definition of Transfer Students
- Importance of Studying Transfer Students
- Survival Analysis Vs. Traditional Regression Methods
Key Factors Impacting Transfer student retention and graduation

- Factors found in regression and descriptive analyses
  - Demographic
  - Community College Credential
  - Transferred Credit Hours
  - Transfer Institution Type

- Factors found in survival analysis
  - Demographic
  - Academic Achievement
  - College Experience
RESEARCH QUESTIONS

1. What is the estimated survival rate of transfer students within eight semesters after enrollment?

2. Are there significant differences between the survival rates of the following sub-groups: age, majors, major changing, transfer GPA, transfer credit hours, financial aid, and enrollment status?

3. What are the effects of covariates on transfer students’ drop out?
VARIABLES

- Age at matriculation
- Transfer credits at entry
- GPA at entry
- Matriculation into a STEM field (Y/N)
- Changed major in first year (Y/N)
- Enrollment intensity in term 1 (full-time/part-time)
- Financial aid received in first term (Y/N)
- Gender
- Race/ethnicity (URM) (Y/N)
STUDY POPULATION

11,267 students who entered institution between 2010 and 2017 as new transfer students in Fall or Summer.

24% URM

54% Female

21% enrolled as STEM majors

16.5% changed majors in first year

74% enrolled full-time in first semester

74% received financial aid
## Study Population

<table>
<thead>
<tr>
<th>Variable</th>
<th>Min</th>
<th>Q1</th>
<th>Median</th>
<th>Q3</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age at matriculation</td>
<td>15</td>
<td>20</td>
<td>21</td>
<td>27</td>
<td>72</td>
</tr>
<tr>
<td>Transfer Credits</td>
<td>1</td>
<td>40</td>
<td>57</td>
<td>69</td>
<td>207</td>
</tr>
<tr>
<td>Transfer GPA</td>
<td>1.16</td>
<td>2.75</td>
<td>3.07</td>
<td>3.43</td>
<td>4</td>
</tr>
</tbody>
</table>
OUTCOME VARIABLE: EVENT

- **Student status at end of study**
  - **Graduated or Continuing enrollment** -> 0
  - **Drop out (did not persist)** -> 1

Example:

<table>
<thead>
<tr>
<th>ID</th>
<th>Time (Semester)</th>
<th>Event (Status)</th>
<th>*Note</th>
<th>Other variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>001</td>
<td>2</td>
<td>1</td>
<td>Dropped out</td>
<td>...</td>
</tr>
<tr>
<td>002</td>
<td>5</td>
<td>0</td>
<td>Graduated</td>
<td>...</td>
</tr>
<tr>
<td>003</td>
<td>8</td>
<td>0</td>
<td>Continuing</td>
<td>...</td>
</tr>
<tr>
<td>004</td>
<td>8</td>
<td>1</td>
<td>Dropped out</td>
<td>...</td>
</tr>
<tr>
<td>005</td>
<td>8</td>
<td>0</td>
<td>Graduated</td>
<td>...</td>
</tr>
</tbody>
</table>
RESEARCH QUESTION 1

What is the estimated survival rate of transfer students within eight semesters after enrollment?

METHOD

Kaplan-Meier Model
SURVIVAL FUNCTION AND KAPLAN-MEIER ESTIMATOR

The survival function, $S(t)$ expresses the probability that a subject’s true survival time $T$ will exceed time $t$.

$$S(t) = Pr(T > t)$$

Example: $S(1) = Pr(T > 1)$

The Kaplan-Meier (KM) estimator is a very popular non-parametric method to estimate the survival function $S(t)$.

$$\hat{S}(t) = \prod_{t_i < t} \left(1 - \frac{events_i}{\text{number at risk}_i} \right)$$

Proportion of those at risk that survive time point $t_i$
### RESULTS: KAPLAN-MEIER MODEL

Table of Kaplan-Meier survival function

<table>
<thead>
<tr>
<th>Semester</th>
<th>Estimated survival probability</th>
<th>lower 95% CI</th>
<th>upper 95% CI</th>
<th>Hazard Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.936</td>
<td>0.932</td>
<td>0.941</td>
<td>0.064</td>
</tr>
<tr>
<td>2</td>
<td>0.847</td>
<td>0.840</td>
<td>0.854</td>
<td>0.095</td>
</tr>
<tr>
<td>3</td>
<td>0.810</td>
<td>0.802</td>
<td>0.817</td>
<td>0.044</td>
</tr>
<tr>
<td>4</td>
<td>0.772</td>
<td>0.764</td>
<td>0.780</td>
<td>0.047</td>
</tr>
<tr>
<td>5</td>
<td>0.751</td>
<td>0.743</td>
<td>0.759</td>
<td>0.027</td>
</tr>
<tr>
<td>6</td>
<td>0.724</td>
<td>0.716</td>
<td>0.733</td>
<td>0.036</td>
</tr>
<tr>
<td>7</td>
<td>0.695</td>
<td>0.686</td>
<td>0.705</td>
<td>0.040</td>
</tr>
<tr>
<td>8</td>
<td>0.681</td>
<td>0.671</td>
<td>0.692</td>
<td>0.020</td>
</tr>
</tbody>
</table>

Estimated survival probability: \( \hat{S}(t) \)

Hazard Rate: \( \frac{events_i}{number\ at\ risk_i} \)
RESEARCH QUESTION 2

Are there significant differences between the survival rates of the following sub-groups: age, majors, major changing, transfer GPA, transfer credit hours, financial aid, gender, race/ethnicity and enrollment status?

METHOD

Stratified Kaplan–Meier Model
Age categories:

- 24 and younger
- 25 and older

Transfer Hours categories:

- 0-29
- 30-59
- 60-89
- 90 or more

Transfer GPA categories:

- 2.49 and below
- 2.50 to 2.99
- 3.00 to 4.00
STRATIFIED BY AGE GROUP

Gehan-Wilcoxon test

<table>
<thead>
<tr>
<th>Age Group</th>
<th>N</th>
<th>Observed</th>
<th>Expected</th>
<th>(O-E)^2/E</th>
<th>(O-E)^2/V</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age Group=1</td>
<td>7727</td>
<td>1665</td>
<td>1908</td>
<td>30.9</td>
<td>121</td>
</tr>
<tr>
<td>Age Group=2</td>
<td>3540</td>
<td>1076</td>
<td>833</td>
<td>70.8</td>
<td>121</td>
</tr>
</tbody>
</table>

Chisq = 121 on 1 degrees of freedom, p = <2e-16

Age Group 1: 24 and younger
Age Group 2: 25 and older
STRATIFIED BY MAJOR CHANGE

Gehan-Wilcoxon test

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Observed</th>
<th>Expected</th>
<th>(O-E)^2/E</th>
<th>(O-E)^2/V</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHANGE D_MAJOR R=N</td>
<td>9405</td>
<td>2379</td>
<td>2248</td>
<td>7.54</td>
<td>50.3</td>
</tr>
<tr>
<td>CHANGE D_MAJOR R=Y</td>
<td>1862</td>
<td>363</td>
<td>493</td>
<td>34.36</td>
<td>50.3</td>
</tr>
</tbody>
</table>

Chisq = 50.3 on 1 degrees of freedom, p = 1e-12
### Gehan-Wilcoxon test

<table>
<thead>
<tr>
<th>Strata</th>
<th>N</th>
<th>Observed</th>
<th>Expected</th>
<th>((O-E)^2/E)</th>
<th>((O-E)^2/V)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FIN_AID_RECEIVED = N</td>
<td>2879</td>
<td>712</td>
<td>688</td>
<td>0.773</td>
<td>1.23</td>
</tr>
<tr>
<td>FIN_AID_RECEIVED = Y</td>
<td>8388</td>
<td>2030</td>
<td>2053</td>
<td>0.259</td>
<td>1.23</td>
</tr>
</tbody>
</table>

Chisq = 1.2 on 1 degrees of freedom, \(p = 0.3\)

![Graph showing survival probability stratified by financial aid received]
STRATIFIED BY FULL TIME / PART TIME

Gehan-Wilcoxon test

<table>
<thead>
<tr>
<th>FT_PT_Flag</th>
<th>N</th>
<th>Observed</th>
<th>Expected</th>
<th>(O-E)^2/E</th>
<th>(O-E)^2/V</th>
</tr>
</thead>
<tbody>
<tr>
<td>FT</td>
<td>8341</td>
<td>1737</td>
<td>2055</td>
<td>49.3</td>
<td>235</td>
</tr>
<tr>
<td>PT</td>
<td>2926</td>
<td>1005</td>
<td>687</td>
<td>147.6</td>
<td>235</td>
</tr>
</tbody>
</table>

Chisq = 235 on 1 degrees of freedom, p = <2e-16
STRATIFIED BY STEM

Gehan-Wilcoxon test

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Observed</th>
<th>Expected</th>
<th>(O-E)^2/E</th>
<th>(O-E)^2/V</th>
</tr>
</thead>
<tbody>
<tr>
<td>STEM_MAJOR=N</td>
<td>8861</td>
<td>2487</td>
<td>2389</td>
<td>4.02</td>
<td>19.1</td>
</tr>
<tr>
<td>STEM_MAJOR=Y</td>
<td>2406</td>
<td>593</td>
<td>691</td>
<td>13.90</td>
<td>19.1</td>
</tr>
</tbody>
</table>

\( \text{Chisq} = 19.1 \) on 1 degrees of freedom, \( p = 1e-05 \)
## Gehan-Wilcoxon test

<table>
<thead>
<tr>
<th>Strata</th>
<th>N</th>
<th>Observed</th>
<th>Expected</th>
<th>((O-E)^2/E)</th>
<th>((O-E)^2/V)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TRCREDIT_Group=1</td>
<td>1216</td>
<td>352</td>
<td>310</td>
<td>5.48</td>
<td>7.434</td>
</tr>
<tr>
<td>TRCREDIT_Group=2</td>
<td>4846</td>
<td>1242</td>
<td>1196</td>
<td>1.75</td>
<td>3.708</td>
</tr>
<tr>
<td>TRCREDIT_Group=3</td>
<td>4371</td>
<td>955</td>
<td>1048</td>
<td>8.17</td>
<td>15.764</td>
</tr>
<tr>
<td>TRCREDIT_Group=4</td>
<td>834</td>
<td>193</td>
<td>187</td>
<td>0.16</td>
<td>0.204</td>
</tr>
</tbody>
</table>

Chisq = 18.6 on 3 degrees of freedom, p = 3e-04

Transfer Credit Group = 1: 0-29
Transfer Credit Group = 2: 30-59
Transfer Credit Group = 3: 60-89
Transfer Credit Group = 4: 90 or more
STRATIFIED BY TRANSFER GPA GROUP

Gehan-Wilcoxon test

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Observed</th>
<th>Expected</th>
<th>(O-E)^2/E</th>
<th>(O-E)^2/V</th>
</tr>
</thead>
<tbody>
<tr>
<td>TRGPA_Group=1</td>
<td>747</td>
<td>219</td>
<td>180</td>
<td>8.18</td>
<td>10.4</td>
</tr>
<tr>
<td>TRGPA_Group=2</td>
<td>419</td>
<td>1108</td>
<td>1014</td>
<td>8.76</td>
<td>16.6</td>
</tr>
<tr>
<td>TRGPA_Group=3</td>
<td>632</td>
<td>1415</td>
<td>1547</td>
<td>11.37</td>
<td>31.1</td>
</tr>
</tbody>
</table>

Chisq = 33.7 on 2 degrees of freedom, p = 5e-08

Transfer GPA Group = 1: 2.49 and below
Transfer GPA Group = 2: 2.50 to 2.99
Transfer GPA Group = 3: 3.00 to 4.00
**STRATIFIED BY GENDER**

### Gehan-Wilcoxon test

<table>
<thead>
<tr>
<th>GENDE R=F</th>
<th>N</th>
<th>Observed</th>
<th>Expected</th>
<th>((O-E)^2/E)</th>
<th>((O-E)^2/V)</th>
</tr>
</thead>
<tbody>
<tr>
<td>6134</td>
<td>1499</td>
<td>1480</td>
<td>0.258</td>
<td>0.669</td>
<td></td>
</tr>
<tr>
<td>GENDE R=M</td>
<td>5132</td>
<td>1242</td>
<td>1262</td>
<td>0.303</td>
<td>0.669</td>
</tr>
</tbody>
</table>

Chisq = 0.7 on 1 degrees of freedom, p = 0.4
STRATIFIED BY RACE/ETHNICITY

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Observed</th>
<th>Expected</th>
<th>(O-E)^2/E</th>
<th>(O-E)^2/V</th>
</tr>
</thead>
<tbody>
<tr>
<td>URM=N</td>
<td>8585</td>
<td>2015</td>
<td>2088</td>
<td>2.59</td>
<td>13</td>
</tr>
<tr>
<td>URM=Y</td>
<td>2681</td>
<td>727</td>
<td>653</td>
<td>8.29</td>
<td>13</td>
</tr>
</tbody>
</table>

Chisq= 13 on 1 degrees of freedom, p= 3e-04

Gehan-Wilcoxon test
FACTORS THAT AFFECT THE OUTCOME

• Age
• Major Change
• Full time / Part time
• STEM / Non-STEM
• Transfer Credit
• Transfer GPA
• Race/Ethnicity
RESEARCH QUESTION 3

What are the effects of covariates on transfer students’ drop out?

METHOD

Cox Proportional Hazards Model

Refined model: Stratified Cox Proportional Hazards Model
COX-PROPORTIONAL HAZARD MODEL

Cox proportional hazards model estimates changes to the hazard function, \( h(t) \). The Cox model can estimate the effects of multiple predictors (covariates) on the hazard function.

\[
h(t|X_1 = x_1) = h_0(t) \cdot e^{(b_1 \cdot x_1)}
\]

- \( h(t|X_1 = x_1) \): the hazard at time \( t \) for a subject with predictor \( X_1 \) equal to the value \( x_1 \)
- \( h_0(t) \): the baseline hazard at time \( t \), the hazard for a subject with all predictors equal to zero
- \( e^{(b_1 \cdot x_1)} \): the hazard ratio comparing the hazard for a subject with \( X_1 = x_1 \) to a subject with \( X_1 = 0 \)

Cox model does not require specification of the baseline hazard function, \( h_0(t) \), the hazard function for a subject with zero on all covariates.
HAZARD RATIO

\[ h(t|X_1 = 0) = h_0(t) \cdot e^{(b_1 \cdot 0)} = h_0(t) \]

\[ h(t|X_1 = 1) = h_0(t) \cdot e^{(b_1 \cdot 1)} \]

Hazard Ratio = \( \frac{h(t|X_1 = 1)}{h(t|X_1 = 0)} = e^{b_1} \)

comparing the hazard for treatment to controls

HR = 0.5 means that treatment has half the hazard of control, or 50% decrease.

HR = 2 means that treatment has double the hazard of control, or 100% increase.

\( e^{b_1} \) express the hazard ratio for a 1-unit increase in the covariate.

\( b_1 \) itself is the log-hazard ratio.
COX MODEL WITH MULTIPLE PREDICTORS

\[ h(t|X_1, X_2, ... X_n) = h_0(t) * e^{(b_1X_1+b_2X_2+\cdots+b_nX_n)} \]

- \( b_i \) : the hazard ratio of covariate \( X_i \)

| Predictor                     | coef | exp(coef) | se(coef) | z     | Pr(>|z|) |
|-------------------------------|------|-----------|----------|-------|----------|
| AGE_AT_MATRIC                 | 0.0132528 | 1.0133410 | 0.0024818 | 5.340 | 9.30e-08 *** |
| STEM MAJORITY                 | -0.2066710 | 0.8132872 | 0.0461482 | -4.478 | 7.52e-06 *** |
| CHANGED MAJORITY              | -0.3293599 | 0.7193840 | 0.0528970 | -6.226 | 4.77e-10 *** |
| TRANSFER UG GPA               | -0.2091169 | 0.8113004 | 0.0285073 | -7.336 | 2.21e-13 *** |
| TRANSFER CREDIT HOURS         | -0.0070707 | 0.9929542 | 0.0009005 | -7.852 | 4.10e-15 *** |
| URMY                          | 0.1628344 | 1.1768417 | 0.0411060 | 3.961 | 7.45e-05 *** |
| FT_PT_FLAGPT                  | 0.5290839 | 1.6973765 | 0.0450434 | 11.746 | < 2e-16 *** |

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
RESULTS: COX-PROPORTIONAL HAZARD MODEL
The standard Cox model assumes proportional hazards, which means that the effects of covariates are constant over time.
CHECKING THE PROPORTIONALITY ASSUMPTION

Smoothed Schoenfeld residuals

Proportional  Time-varying  Time-varying
CHECKING THE PROPORTIONALITY ASSUMPTION

The `cox.zph` function will test proportionality of all the predictors in the model. A p-value less than 0.05 indicates a violation of the proportionality assumption.

<table>
<thead>
<tr>
<th>Predictor</th>
<th>chisq</th>
<th>df</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGE_AT_MATRIC</td>
<td>0.608</td>
<td>1</td>
<td>0.436</td>
</tr>
<tr>
<td>STEM_MAJOR</td>
<td>42.103</td>
<td>1</td>
<td>8.7e-11</td>
</tr>
<tr>
<td>CHANGED_MAJOR</td>
<td>129.063</td>
<td>1</td>
<td>&lt; 2e-16</td>
</tr>
<tr>
<td>TRANSFER_UG_GPA</td>
<td>7.537</td>
<td>1</td>
<td>0.006</td>
</tr>
<tr>
<td>TRANSFER CREDIT HOURS</td>
<td>4.197</td>
<td>1</td>
<td>0.041</td>
</tr>
<tr>
<td>URM</td>
<td>1.813</td>
<td>1</td>
<td>0.178</td>
</tr>
<tr>
<td>FT_PT_Flag</td>
<td>40.799</td>
<td>1</td>
<td>1.7e-10</td>
</tr>
</tbody>
</table>

violation of the proportionality assumption
In the stratified Cox model:

- the Cox model is estimated separately in each stratum
- the baseline hazard function is allowed to be different across strata - this can accommodate the non-proportional effects of the stratification variable
- the parameter estimates are then averaged across strata to generate one final set of estimates
DEAL WITH TIME-VARYING COEFFICIENTS: STRATIFIED COX MODEL

Stratum 1: Semester 1, 2  
Stratum 2: Semester 3, 4  
Stratum 3: Semester 5 - 8

| Variable                        | Coef  | Exp(coef) | Se(coef) | Z     | Pr(>|z|) |
|--------------------------------|-------|-----------|----------|-------|----------|
| AGE_AT_MATRIC                  | 0.0133526 | 1.0134421 | 0.0024793 | 5.386 | 7.22e-08 *** |
| TRANSFER_UG_GPA                | -0.2083737 | 0.8119036 | 0.0285816 | -7.290 | 3.09e-13 *** |
| TRANSFER_CREDIT_HOURS         | -0.0070348 | 0.9929899 | 0.0008987 | -7.827 | 4.98e-15 *** |
| URMY                           | 0.1612789 | 1.1750127 | 0.0411088 | 3.923 | 8.74e-05 *** |
| STEM_MAJOR:strata(tgroup)tgroup=1 | -0.5063585 | 0.6026863 | 0.0672493 | -7.530 | 5.09e-14 *** |
| STEM_MAJOR:strata(tgroup)tgroup=2 | 0.0080230 | 1.0080553 | 0.0854077 | 0.094 | 0.92516 |
| STEM_MAJOR:strata(tgroup)tgroup=3 | 0.2878625 | 1.3335740 | 0.0976119 | 2.949 | 0.00319 ** |
| strata(tgroup)tgroup=1:CHANGED_MAJOR | -1.1301415 | 0.3229875 | 0.0976580 | -11.572 | < 2e-16 *** |
| strata(tgroup)tgroup=2:CHANGED_MAJOR | 0.3989203 | 1.4902148 | 0.0818070 | 4.876 | 1.08e-06 *** |
| strata(tgroup)tgroup=3:CHANGED_MAJOR | 0.0216448 | 1.0218807 | 0.1094653 | 0.198 | 0.84325 |
| strata(tgroup)tgroup=1:FT_PT_Flag | 0.6555032 | 1.9261116 | 0.0549751 | 11.924 | < 2e-16 *** |
| strata(tgroup)tgroup=2:FT_PT_Flag | 0.4972412 | 1.6441791 | 0.0785099 | 6.333 | 2.40e-10 *** |
| strata(tgroup)tgroup=3:FT_PT_Flag | 0.1225132 | 1.1303340 | 0.1001945 | 1.223 | 0.22142 |

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
CONCLUSIONS

• With each additional year of age at matriculation, the drop out probability increases 1% on average.

• With each 1-point increase in transfer UG GPA, the drop out probability decreases 19% on average.

• With each 10 transfer credits increase, the drop out probability decreases 7% on average.

• URM students are 18% more likely to drop out compared with non-URM students on average.
In the first two semesters, STEM majors are 40% less likely to drop out than non-STEM students, but after 4 semesters, they are 33% more likely to drop out than non-STEM students.

In the first two semesters, students who changed major are 68% less likely to drop out than those did not change major. But from semester three to four, students who changed major are 50% more likely to drop out than those that did not change major. After four semesters, there is no difference.

In the first two semesters, part-time students are 90% more likely to drop out than full-time students. From semester three to four, they are 60% more likely to drop out. After four semesters, there is no difference.
DISCUSSION: BENEFITS OF SURVIVAL ANALYSIS

• **ALLOWS US TO MEASURE THE EFFECT OF FACTORS IN STUDENT SUCCESS THAT VARY OVER TIME THAT LOGISTIC REGRESSION OR MULTIPLE LINEAR REGRESSION ARE UNABLE TO DO.**

• **PROVIDES INSIGHT INTO WHEN AND HOW INTERVENTIONS MIGHT BE MOST SUCCESSFUL, E.G.,**
  
  • HELPING STUDENTS TO CHOOSE THE APPROPRIATE MAJOR IN THE FIRST YEAR
  
  • PROVIDING RESOURCES TO HELP TRANSFER STUDENTS ENROLL FULL-TIME, AT LEAST IN THEIR FIRST FEW SEMESTERS
  
  • REVIEWING POLICIES & PRACTICES THAT MIGHT PREVENT STUDENTS FROM GETTING FULL CREDIT FOR TRANSFERRED COURSES THAT COUNT TOWARD A DEGREE PROGRAM AND/OR GRADUATION
CONTACT INFORMATION

• Beverly King  
  kingb14@ecu.edu

• Margot Neverett  
  neverettm@ecu.edu

• Franklin Zhou  
  zhou21@ecu.edu

• Yihui Li  
  liy17@ecu.edu

• Kyle Chapman  
  chapmank@ecu.edu
Thank you for attending the 2023 NCAIR Annual Conference!

There’s a QR code in your program for a conference evaluation form. You’ll also get an e-mail following the conference with a link to the form, which will be available until 4/18.

Please take the opportunity at your earliest convenience to let the planning committee know your thoughts about this year’s conference and where you would like to meet next year.
Appendix: R Code

# Loading packages  
library(survival)  
library(tidyverse)  
library(survminer)  
library(broom)  

# Data importing and processing  
dataset <- readxl::read_xlsx("Survival_Analysis_dataset.xlsx", sheet = "Sheet1") %>% data.frame()  

# Subgroup Data  
dataset$Age_Group <- with(dataset, case_when (AGE_AT_MATRIC <= 24 ~ 1,  
                                 AGE_AT_MATRIC > 24 ~ 2))  

dataset$TRGPA_Group <- with(dataset, case_when (TRANSFER_UG_GPA < 2.5 ~ 1,  
                                           TRANSFER_UG_GPA < 3 ~ 2,  
                                           TRANSFER_UG_GPA <= 4 ~ 3))  

dataset$TRCREDIT_Group <- with(dataset, case_when (TRANSFER_CREDIT_HOURS < 30~ 1,  
                                          TRANSFER_CREDIT_HOURS < 60 ~ 2,  
                                          TRANSFER_CREDIT_HOURS < 90 ~ 3,  
                                          TRANSFER_CREDIT_HOURS >= 90 ~ 4))  

dataset <- dataset %>%  
mutate(across(c(NCCCS_TRANSFER_IND, UNC_TRANSFER_IND, CHANGED_MAJOR,  
               FIN_AID_RECEIVED, NEED_BASED, MERIT, PELL, LOAN, FT_PT_Flag,  
               COURSE_DELIVERY, STEM_MAJOR, Age_Group, TRGPA_Group, TRCREDIT_Group, GENDER,  
               URM), as.factor)) %>%
select(Time, Event_Q, AGE_AT_MATRIC, TRANSFER_CREDIT_HOURS, TRANSFER_UG_GPA, CHANGED_MAJOR, FIN_AID_RECEIVED, FT_PT_Flag, STEM_MAJOR, Age_Group, TRGPA_Group, TRCREDIT_Group, GENDER, URM)

attach(dataset)

# fit the Kaplan-Meier model for the entire data --------------------
km.model.Q <- survfit(Surv(Time,Event_Q)~1, type="kaplan-meier")

# summary the "step function"
summary(km.model.Q)

# plot the KM function
plot(km.model.Q,conf.int = T, xlab = "semester", ylab = "Survival probility", main= "Kaplan Meier Model of Drop Out",las=1, mark.time = T)

# Stratified KM Model ------------------------------------------

# variable PT/FT
km.model.Q.PT_FT <- survfit(Surv(Time,Event_Q)~FT_PT_Flag, type="kaplan-meier")

summary(km.model.Q.PT_FT)

## use ggsurvplot to plot

ggsurvplot(km.model.Q.PT_FT,data = dataset, conf.int=T, xlab = "Semester", mark.time = F, censor=F)

## gehan-wilcoxon test: set rho = 1

## H0: survival curves across 2 or more groups are equivalent

## HA: survival curves across 2 or more groups are not equivalent

survdiff(Surv(Time,Event_Q)~FT_PT_Flag, rho = 1)

# Cox Proportional Hazard Model ----------------------------------
# fit the Cox-proportional hazard model

cox.model <- coxph(Surv(Time, Event_Q) ~ AGE_AT_MATRIC + STEM_MAJOR +
                  CHANGED_MAJOR + TRANSFER_UG_GPA + TRANSFER_CREDIT_HOURS + URM +
                  FT_PT_Flag, data = dataset)

summary(cox.model)

## Checking proportional hazard assumption

# H0: covariate effect (Hazards) is constant (proportional) over time
# Ha: covariate effect (Hazards) changes over

zp <- cox.zph(cox.model)

plot(cox.zph(cox.model)[1], xlab = "semester")

abline(h=0,col="red")

## stratified cox ph model

# stratified cox ph model

### set cut point of semesters

dataset.split <- survSplit(Surv(Time, Event_Q)~, data=dataset, cut = c(2, 4), episode = "tgroup", id="id")

dataset.split$FT_PT_Flag <- as.numeric (dataset.split$FT_PT_Flag)

dataset.split$STEM_MAJOR <- as.numeric (dataset.split$STEM_MAJOR)

dataset.split$CHANGED_MAJOR <- as.numeric (dataset.split$CHANGED_MAJOR)

## stratified model

stratified_model1 <- coxph(Surv(tstart, Time, Event_Q) ~ AGE_AT_MATRIC + STEM_MAJOR:strata(tgroup) +
                           CHANGED_MAJOR:strata(tgroup) + TRANSFER_UG_GPA + TRANSFER_CREDIT_HOURS + URM +
                           FT_PT_Flag:strata(tgroup), data=dataset.split)

summary(stratified_model1)

cox.zph(stratified_model1)