Analysis on factors associated with time to death: Scottish ISD Study

1. Introduction

In this study, we analyse the relations of age, gender, comorbidities, receiving an operation, length of stay, and hospital costs to time to death.

Research has found comorbidities have a negative impact on survival time in gastric, colorectal, and lung cancer patients (Morishima et al, 2019). Another study found myocardial infarction, renal disease, and metastatic carcinoma to be the most common comorbidities to reduce survival time in hospital patients of any indication (Anderson et al, 2019). Research has found high healthcare costs to be associated with an increased risk of mortality (Van der Ree et al, 2019). Patients with longer lengths of stay are associated with higher mortality risks (Kim et al, 2019). Operations have been associated with reduced survival time in patients with a variety of indications, including ovarian cancer, breast cancer, cholangiocarcinoma, among other indications (Marsh et al, 1992, de Castro et al, 2022). Age has been found to be associated with a decrease in predicted survival time (Versteegh et al, 1992). Studies release inconsistent results about whether gender is associated with survival time: some find survival times are indistinguishable across gender (Kravdal et al, 2016), some find females have a longer survival time (Gagnon et al, 2017), and some find males have a longer survival time (Crimmins et al, 2016). Previous research has leveraged a variety of survival analysis methods, including multivariate cox proportional hazards models, multivariate logistic regression, multivariate linear regression models, exponential, weibull, generalised-gamma, etc. (Chugtai et al, 2018; Chan et al, 2016; Meara et al, 2015; Sun et al, 2014)

While there are many studies that examine factors associated with time to death, many of them are indication-specific. In this study, we further analyse the effects on a large volume of hospital patients who have been admitted for various indications.

2. Data

2.1 Description of Data

The dataset used in our analysis comes from the Scottish ISD, containing information on hospital admissions for inpatients from the General Medicine Speciality.

Time to death (“*time*”) is our dependent variable, and variables length of stay (“*los*”), cost of stay (“*costofstay*”), age (“*ageyrs*”), sex (“*female*”), comorbidity (“*comorb*”), operation (“*oper*”) are our covariates of interest. Table 1 provides a summary of the variables used in our analysis. Each patient has a unique *patientid* that is associated with their data.

Table 1: Overview of variables of interest

```
 Observations:        14,490
    Variables:            13                  17 Mar 2023 08:47
```

Variable name	Storage type	Display format	Value label	Variable label
patientid	str8	%9s		**Patient id**
mineadmdate	int	%td		**Date of admission to hospital**
edeathdate	int	%td		**Date of death**
elastdate	int	%td		**Last date observed in study**
los	float	%9.0g		**Hospital length of stay in days**
costofstay	float	%9.0g		**Total cost upto death or censoring: costperday*lengthofstay**
ageyrs	float	%9.0g		**Age in years**
female	float	%9.0g	sex	**Sex**
agedeathyrs	float	%9.0g		**Age at death in years**
comorb	float	%9.0g	yesno	**Has comorbidities**
oper	float	%9.0g	yesno	**Had an operation**
death	float	%9.0g	yesno	**Died**
y	float	%9.0g		**Total cost upto death or censoring: costperday*lengthofstay**

```
Sorted by:
     Note: Dataset has changed since last saved.
```

Table 2: Descriptive statistics

Variable	Obs	Mean	Std. dev.	Min	Max
patientid	0				
mineadmdate	14,490	10570.32	1654.395	7671	13601
edeathdate	10,765	12360.36	2149.101	7677	16159
elastdate	14,490	16070	0	16070	16070
los	14,490	12.9873	25.05514	.5	490
costofstay	14,490	1322.107	2550.613	50.9	49882
ageyrs	14,490	65.95121	11.3653	21.13895	100.7283
female	14,490	.4799862	.4996165	0	1
agedeathyrs	10,765	73.96935	9.870243	30.99247	103.0938
comorb	14,490	.7218081	.4481239	0	1
oper	14,490	.0645963	.2458206	0	1
death	14,490	.7429262	.4370355	0	1
y	14,490	2743.689	2121.263	.5	8397
x	14,490	1322.107	2550.613	50.9	49882
lnx	14,490	6.502699	1.151484	3.929863	10.81742
time	14,490	2743.689	2121.263	.5	8397

2.2. Missing Data, Skewness, Heteroskedasticity, and Censoring

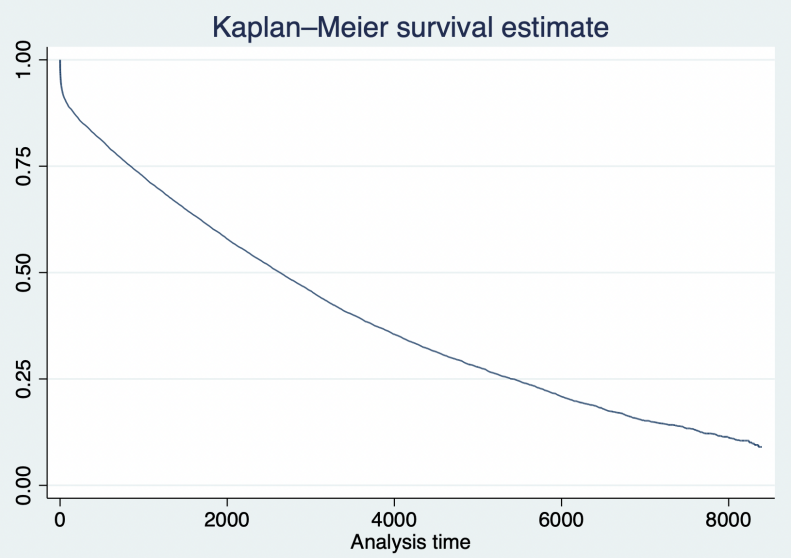
As shown in Table 2, our data contains 14,490 observations. As there is no missing data, no adjustments are needed to account for missing values. A non-skewed model assumes that variance in a particular variable remains constant overtime, which is often necessary for modelling data. When summarising the cost per day variable, the skewness is found to be high (7.61, Appendix Table 6.1), suggesting the data is heavily right skewed. That is, there exists a small proportion of patients that pay really high costs. In order to account for this skewness, the log of cost is computed and included in future model experiments. Table 6.1 summarises this new variable, showing the skewness decreases to -0.10, normalising it for

our analysis (Appendix Table 6.1). The skewness of our dependent variable “time to death” is low (0.52, Appendix Table 6.2), suggesting the distribution is relatively symmetric.

Heteroskedasticy occurs when there is an unequal distribution in the variance of residuals (Robinson, 1987). In our test for heteroskedasticity for models with untransformed and log costs, both graphs show a bit of heteroskedasticity, given their curve and the change in spread of the data points (Graph 6.4, 6.5). This might invalidate our data, and suggests GLM or exponential might be a good fit to account for this (over OLS).

Given our data is collected in a specific time frame, death is observed in only a portion of the patients. The variable “*death*” indicates whether that particular patient died over the period of the data collection. When this variable is not equal to 1, it suggests the observation is censored, as death did not occur for that individual. Further analysis into censored data suggests that death is observed in 10,725 patients, and not observed in 3,725 patients (Appendix, Table 6.5). After estimating the means, we find that the largest observed analysis time is censored, meaning our mean survival time (3,295.61 days(*)) is underestimated in our analysis (Appendix, Table 6.6). We find the extended mean to be 3,610.87 days, which estimates the mean while accounting for censoring through the utilisation of a Kaplan-Meier function. When analysing the “time to death” variable in more depth, we find the mean time for all observations to be 2,743.69 days, 2,030.80 days for uncensored, and 4,803.89 days for censored observations. The standard deviations are higher than that of a normal distribution, ranging from 1,545.14 - 2,121.26.

Graph 1: Kaplan-Meier Survival Curve to Account for Censoring



3. Methods

3.1 Model Overview, Assumptions

There are several different models used in survival analysis, including the Cox proportional hazards model, the Weibull model, and the Exponential model. The Weibull and Exponential models are parametric, meaning we are assuming we are drawing the data from a particular distribution. When the distribution for our dependent variable (survival time) is unknown, we can estimate our coefficients via a semiparametric model, such as the cox proportional hazards model. The Weibull model assumes a linear change in hazard rate. The exponential model assumes the hazard rate is constant. While the Cox model allows hazard rate to fluctuate, its validity is contingent on the proportional hazards assumption, which states that the hazard rates between each covariate must remain constant. It also assumes that each covariate has a linear relationship with log hazard. The cox assumptions can be tested through Kaplan Meier estimates and residual plots.

All models assume independence in the survival times across individuals, and that the probability of survival is independent of the time that has progressed since the start of the study (Lindsey et al, 1998). Lastly, we are assuming that the incidence of censoring is independent from the chance of the development of the event of interest (time to death).

3.2 Testing Models, Strengths and Weaknesses

In answering whether to use a parametric or semi parametric model, we need to consider whether we can assume the distribution of our model is known. Given the challenges in doing this, a Weibull and exponential model were both run and the log-likelihood output was analysed to determine goodness-of-fit. An Akaike Information Criterion (AIC) was run for Weibull and Exponential models to further compare the prediction errors. Furthermore, a Cox proportional hazards model was run to compare the outputs and test the strength of the model. Kaplan meier survival curves were run by each covariate to test whether hazard rates were parallel. Log-log plots were run to test the proportional hazard assumptions for Cox. If the assumptions regarding the underlying distribution are correct, a parametric model might be more robust than Cox. Weibull tends to be more robust than exponential, as it allows for more variation in the fitting of the model. While all models can evaluate several covariates, Cox is the only model that is able to examine time-to-event data with respect to several variables at the same time. If we are able to correctly make assumptions about the underlying distribution, parametrics models tend to be more robust than semiparametric models. However, if these assumptions are not correct, results from parametric models are at risk of being biassed.

3.3 Equations

The equations for each prospective model are displayed below.

Equation for cox proportional hazards model:

$\lambda(t \mid x_{i1}, x_{i2}, \dots, x_{ip}) = \lambda_0(t)\exp(\mathbf{\Sigma} x_{ij}, \beta_j)$

$\lambda(t \mid x_{i1}, x_{i2}, \dots, x_{ip})$ = hazard at time t

$x_{i1}, x_{i2}, \dots, x_{ip}$ = covariates

$\lambda_0(t)$ = an arbitrary nonnegative function
B_j, = the impact of each covariate on the hazard ratio

Equation for Weibull:
$\log(h(t|x)) = \log(\lambda(x)) + (\lambda(x)-1) * \log(t/\theta(x))$

- h(t) = the hazard function
- $x_{i1}, x_{i2}, \dots, x_{ip}$ = covariates
- t = time
- λ = shape parameter
- Θ = scale parameter, which determines the time scale of the hazard function

Equation for Exponential model
$\log(\lambda) = \boldsymbol{\alpha} + \beta_1 x_1 + \dots + \beta_k x_k$

- λ = parameter (that is fixed but unknown)
- $x_1, \dots, x_k$ = covariates
- $e^{\beta \Delta x}$ = estimate of the relative hazard

4. Results

Table 3: Goodness-of-fit metrics

	Log-likelihood	AIC
Exponential with untransformed costs	-25,289.41	50,590.81
Exponential with transformed costs	-25,285.79	50,585.58
Weibull with untransformed costs	-24,674.12	49,362.24
Weibull with transformed costs	-24,671.02	49,358.85
Cox with untransformed costs	-93,544.68	N/A
Cox with transformed costs	-93,542.68	N/A
Stratified cox with untransformed costs	-3,631.51	N/A
Stratified cox with transformed costs	-3,636.05	N/A

In our initial goodness-of-fit tests, we look at the log likelihood to see which model might be the best fit to our data. For our parametric models (Exponential and Weibull), we use AIC outputs to determine which parametric model is the best fit. Models with smaller AIC fit the data better (Lindsey et al, 1998). We find that the Weibull model with transformed costs

(ln(costs)) fits the data best, given it has the lowest AIC output (49,358.85). However, the very negative log likelihoods of these parametric models suggests these might not be the best fit to our data (Bosman et al, 2000). Therefore, we also look at the semi-parametric Cox proportional hazards model. The log-likehoods for our Cox models are closer to zero, suggesting the Cox model might be a better fit for our data. Within our different Cox models, we find that the Cox stratified model with untransformed costs is the best fit, given it has the higher log likelihood value. However, given this model omits cost from the output, the stratified Cox with log transformed costs would give us more information on our variables of interest, for only a marginally lower log-likelihood. Further analyses are conducted to test the proportional hazards assumptions and fit of the cox proportional hazards model to our data.

Testing Assumptions:

Table 4: Tests of proportional hazard model assumptions for unstratified cox model

```
. estat phtest, detail

Test of proportional-hazards assumption

Time function: Analysis time
```

	rho	chi2	df	Prob>chi2
ageyrs	-0.06065	39.16	1	0.0000
female	-0.02763	8.24	1	0.0041
comorb	-0.17684	327.82	1	0.0000
oper	0.03693	14.72	1	0.0001
los	-0.02696	6.37	1	0.0116
lnx	0.05523	37.15	1	0.0000
Global test		436.13	6	0.0000

Table 5: Tests of proportional hazard model assumptions for stratified cox model

```
. estat phtest, detail

Test of proportional-hazards assumption

Time function: Analysis time
```

	rho	chi2	df	Prob>chi2
female	-0.00639	0.45	1	0.5036
comorb	-0.05889	37.50	1	0.0000
oper	0.01273	1.77	1	0.1828
los	-0.00932	1.15	1	0.2834
lnx	0.03210	12.15	1	0.0005
Global test		51.62	5	0.0000

.

In our initial test for the proportional hazards model for Cox, we find that the p-values are less than 0.05 for all of our variables of interest, suggesting the proportional hazards model has been violated and our Cox regression results might not be accurate. In addition, the fact that our lines are not parallel in our log-log plots of survival (Appendix, Graphs 4-9) further suggests a violation of the proportional hazards assumption. To adjust for these violations,

we also run a Cox proportional hazards model that is stratified by age. After testing for the proportional hazards assumption for the stratified model, we find that the p-values are now larger for several variables (excluding *comorb* and *ln(costs)*), suggesting this model might be a better fit for our data given assumptions are not violated for most variables of interest. We also ran a time variant cost model; however, given the very low log likelihood (~ -90,000), this model does not seem to be a good fit.

Table 6: Hazard ratios for chosen models

variables	Exponential model hazard ratio	p-value	Cox stratified model hazard ratio	p-value	Weibull model hazard ratio	p-value
ageyrs	1.0537	0.000	N/A	N/A	1.0475	0.000
female	0.9436	0.003	0.9627	0.359	0.9527	0.014
comorb	1.6347	0.000	2.0800	0.000	1.5556	0.000
oper	0.8803	0.003	0.8463	0.068	0.8547	0.000
los	1.0027	0.000	1.0031	0.003	1.0022	0.013
ln(cost)	1.0324	0.008	1.0050	0.835	1.029	0.000

Interpretation of results from chosen model:

In our outputs for various models, a hazard ratio less than one suggests a negative association between that variable and time to death (survival time). A hazard rate greater than one suggests a positive association between that variable and time to death, and a hazard rate equal to one suggests there is no effect of that variable on survival time (Spotswood et al, 2004).

Given the stratified Cox proportional hazards model with transformed costs seems to be the best fit model, we choose to evaluate the hazard ratio of these. The gender (*female*) variable output (0.9627) suggests that being female is associated with a decrease in survival time; however, the high p-value (0.359) suggests this effect is not significant. The comorbidity (*comorb*) hazard ratio (2.0800) suggests that having comorbidities decreases your length of survival. The low p-value (0.000; > 0.05) suggests this effect is significant at the 5% significance level. The operations (*oper*) hazard ratio (0.8463) suggests that having an operation increases one's length of survival. However, the p-value is 0.068, suggesting the results are not statistically significant at the 5% significant level, but are at a slightly higher significant level (i.e. 7%). Our length of stay (*los*) hazard ratio is 1.0031, suggesting that a longer length of stay is associated with a shorter survival time. The p-value of 0.003 suggests this result to be significant at the 5% level. Lastly, our *ln(cost)* hazard ratio is 1.0050, suggesting that patients with higher costs tend to have shorter survival times. The p-value is high (0.835), suggesting these results are not statistically significant.

As this model is stratified by age, all of these effects and their associations with survival time hold true while adjusting for the effect of age. Since age is omitted as a variable of interest in this model, we turn to the coefficients of other models to example whether age is a factor that impacts survival time. For both Exponential and Weibull, we find that the age coefficient is greater than one, suggesting that an increase in age is associated with a decrease in survival time. Both of these coefficients are significant at the 5% level (p-values of 0.000 and 0.000).

Since the proportional hazard tests suggest there might be a risk of bias in our comorbidity and cost coefficients, we look at the magnitudes to evaluate if there are any differences. We find that the hazard ratio for comorbidity is greater than one in all of our models - so, although there might be a risk of bias, our results suggest that comorbidities are associated with a decrease in survival time. We also find that the hazard ratio for *ln(costs)* is greater than one in all models, strengthening our finding that an increase in cost is associated with a decreased survival time. This finding is statistically significant at the 5% significance level for Weibull and Exponential distributions (p-values of 0.000 and 0.008).

5. Discussion

Comparisons in hazard ratios across all models suggests consistently in our findings: having an older age, comorbidities, a longer length of stay, and higher hospital costs are associated with a shorter survival time, whereas undergoing an operation is associated with a longer survival time. The magnitudes of these hazards ratios are also comparable across models. Side by side, the graphs of analysis time and survival look very similar for all models (Appendix Graphs 2, 3, 10, 11). There do not seem to be any significant associations between gender and time to death. Transforming our *cost* variable into *ln(costs)* seems to eliminate skewness and improve the fit of our models (Appendix Table 6.1). This is further proved given the log-likelihood values tend to increase in models involving our transformed variable (Appendix Tables 7-18). Nonetheless, the negative log likelihoods suggest none of our models are a “perfect” fit for our data, although the best fit model appears to be Cox proportional hazards stratified by age (Table 6). The p-values for the majority of variables (*ageyrs, comorb*, and *los*) are statistically significant; however, results for operation (*oper*) and *ln(cost)* variables should be interpreted with caution for our stratified Cox model. There are also some discrepancies in our *female* variable, as our Exponential and Weibull models suggest being female significantly reduces survival time at a 5% significance level. However, given the low values of log likelihoods for this model, these findings might be impacted by misassumptions in the parametric fit of the data, making our stratified cox model findings more robust for this variable. Overall, while there is consistency in our findings and we have found a model that fits relatively well with our data, we should interpret our results with caution given the subtle nuances in the data and their significance. In future research, we could analyse these effects for demographic-specific characteristics. As these results pertain to Scottish hospital data in particular, we could also run comparable analyses on hospital data from other regions to test the generalizability of our findings.

6. Appendix

Table 6.1 Skewness measures for cost variable

Total cost upto death or censoring: costperday*lengthofstay

Percentiles		Smallest		
1%	50.9	50.9		
5%	101.8	50.9		
10%	101.8	50.9	Obs	14,490
25%	407.2	50.9	Sum of wgt.	14,490
50%	712.6		Mean	1322.107
		Largest	Std. dev.	2550.613
75%	1221.6	44995.6		
90%	2545	46013.6	Variance	6505626
95%	4377.4	47846	Skewness	7.621959
99%	12317.8	49882	Kurtosis	87.69754

```
. summarize lnx, detail
```

lnx

Percentiles		Smallest		
1%	3.929863	3.929863		
5%	4.62301	3.929863		
10%	4.62301	3.929863	Obs	14,490
25%	6.009305	3.929863	Sum of wgt.	14,490
50%	6.56892		Mean	6.502699
		Largest	Std. dev.	1.151484
75%	7.107917	10.71432		
90%	7.841886	10.73669	Variance	1.325915
95%	8.384211	10.77574	Skewness	-.0976126
99%	9.4188	10.81742	Kurtosis	3.45791

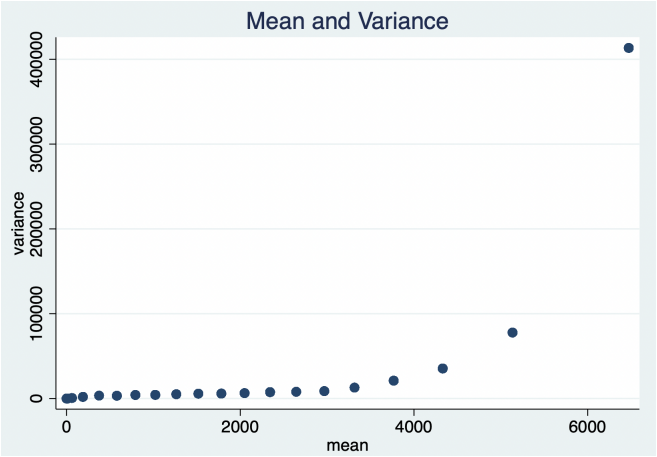
Table 6.2 Skewness measures for time variable

```
. summarize y, detail
```

y

Percentiles		Smallest		
1%	1	.5		
5%	9	.5		
10%	76	.5	Obs	14,490
25%	843	.5	Sum of wgt.	14,490
50%	2588.5		Mean	2743.689
		Largest	Std. dev.	2121.263
75%	4158	8386		
90%	5798.5	8387	Variance	4499756
95%	6714	8391	Skewness	.5194166
99%	7992	8397	Kurtosis	2.418441

Graph 6.3: Mean and variance for costs



Graph 6.4: Mean and variance for ln(costs)

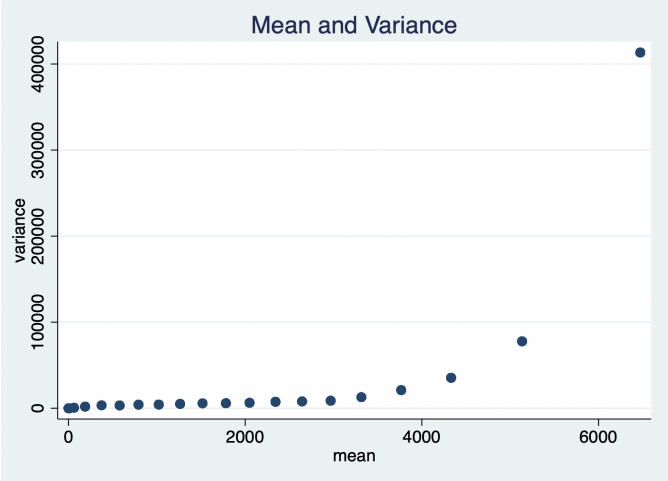


Table 6.5: Censoring data

```
> stset time, failure(death==1)

Survival-time data settings

         Failure event: death==1
Observed time interval: (0, time]
     Exit on or before: failure

------------------------------------------------------------------------------
     14,490  total observations
          0  exclusions
------------------------------------------------------------------------------
     14,490  observations remaining, representing
     10,765  failures in single-record/single-failure data
 39756052.5  total analysis time at risk and under observation
                                                At risk from t =         0
                                     Earliest observed entry t =         0
                                          Last observed exit t =     8,397

. tab death

       Died |      Freq.     Percent        Cum.
------------+-----------------------------------
         No |      3,725       25.71       25.71
        Yes |     10,765       74.29      100.00
------------+-----------------------------------
      Total |     14,490      100.00
```

Table 6.6: Extended and restricted means

```
.
. ///Estimate means and associated confidence intervals of survival time
> stci, rmean

        Failure _d: death==1
  Analysis time _t: time

             |  Number of  Restricted
             |   subjects        mean      Std. err.    [95% conf. interval]
-------------+------------------------------------------------------------
       Total |      14490    3295.619(*)   24.41709      3247.76    3343.48

(*) largest observed analysis time is censored, mean is underestimated

. stci, emean

        Failure _d: death==1
  Analysis time _t: time

             |  Number of    Extended
             |   subjects        mean
-------------+----------------------
       Total |      14490     3610.87

. sts

        Failure _d: death==1
  Analysis time _t: time

.
```

Table 7: Regression output of cox proportional hazards model with untransformed costs

```
.
. stcox ageyrs female comorb oper los x

        Failure _d: death==1
  Analysis time _t: time

note: x omitted because of collinearity.
Iteration 0:   log likelihood = -95735.671
Iteration 1:   log likelihood = -93588.234
Iteration 2:   log likelihood = -93546.286
Iteration 3:   log likelihood = -93544.687
Iteration 4:   log likelihood = -93544.681
Refining estimates:
Iteration 0:   log likelihood = -93544.681

Cox regression with Breslow method for ties

No. of subjects =     14,490                         Number of obs =  14,490
No. of failures =     10,765
Time at risk    = 39,756,053
                                                     LR chi2(5)    = 4381.98
Log likelihood = -93544.681                          Prob > chi2   =  0.0000
```

_t	Haz. ratio	Std. err.	z	P>|z|	[95% conf.	interval]
ageyrs	1.05569	.0010394	55.04	0.000	1.053655	1.057729
female	.9399044	.0185421	-3.14	0.002	.9042562	.9769579
comorb	1.650869	.0375699	22.03	0.000	1.578852	1.726172
oper	.9089513	.039181	-2.21	0.027	.8353125	.989082
los	1.003214	.0003098	10.39	0.000	1.002607	1.003821
x	1	(omitted)				

Graph 2: Cox regression with untransformed costs

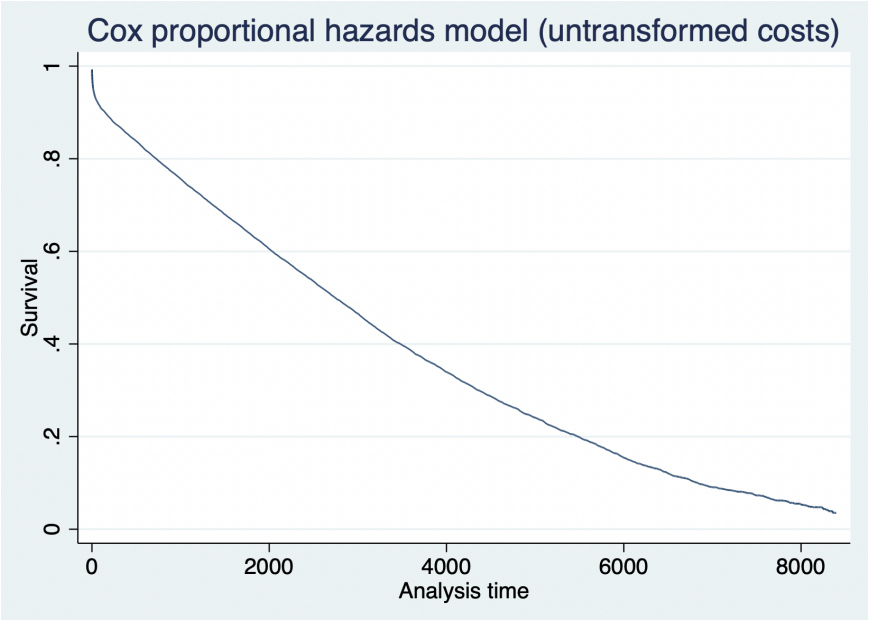
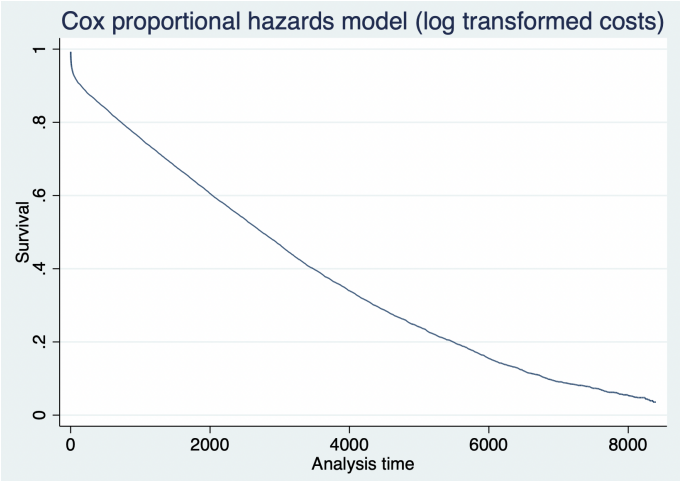


Table 8: Regression output of cox proportional hazards model with log transformation of costs

```
. stcox female comorb oper los lnx

        Failure _d: death==1
  Analysis time _t: time

Iteration 0:   log likelihood = -95735.671
Iteration 1:   log likelihood = -95144.812
Iteration 2:   log likelihood = -95103.283
Iteration 3:   log likelihood = -95100.415
Iteration 4:   log likelihood = -95100.392
Iteration 5:   log likelihood = -95100.392
Refining estimates:
Iteration 0:   log likelihood = -95100.392

Cox regression with Breslow method for ties

No. of subjects =     14,490                     Number of obs =  14,490
No. of failures =     10,765
Time at risk    = 39,756,053
                                                 LR chi2(5)    = 1270.56
Log likelihood = -95100.392                      Prob > chi2   =  0.0000
```

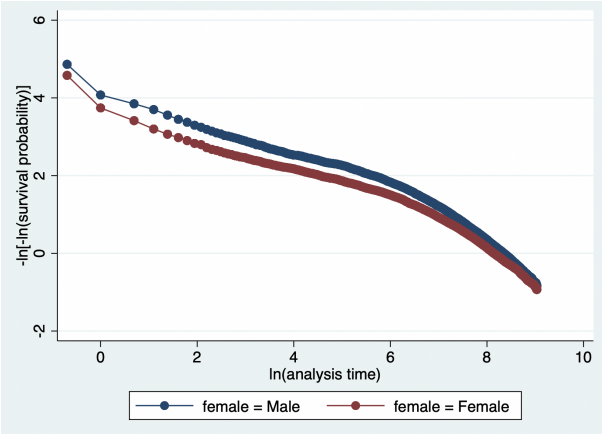
_t	Haz. ratio	Std. err.	z	P>|z|	[95% conf.	interval]
female	1.141503	.022111	6.83	0.000	1.098979	1.185673
comorb	1.81872	.0413784	26.29	0.000	1.739402	1.901656
oper	.8356536	.0360749	-4.16	0.000	.7678567	.9094366
los	1.002122	.0004256	4.99	0.000	1.001288	1.002956
lnx	1.124117	.0133941	9.82	0.000	1.098169	1.150678

Graph 3: Cox regression with log transformed costs

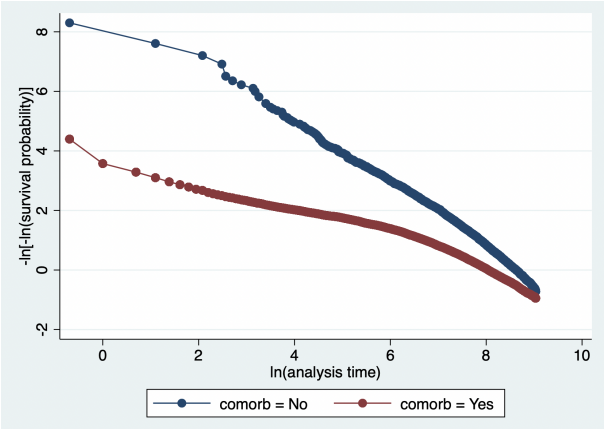


Graphs 4-9: Cox proportional Log log plots of survival:

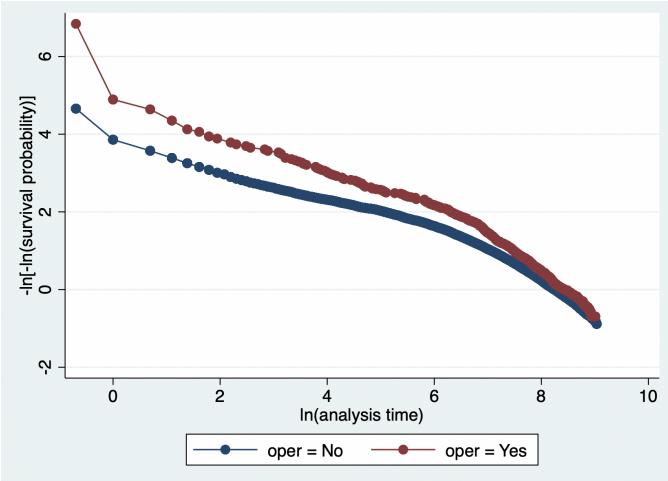
Graph 4: by gender



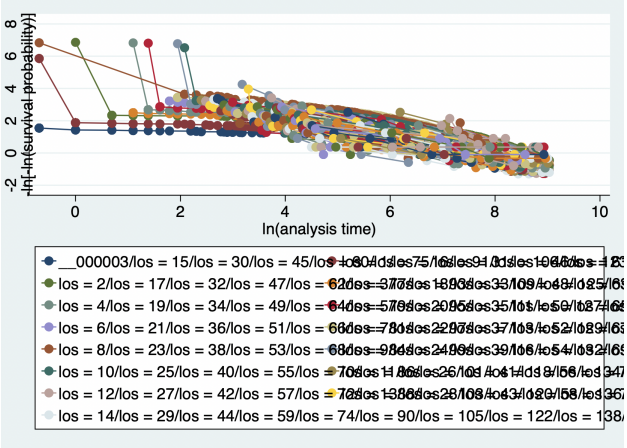
Graph 5: by comorbidities



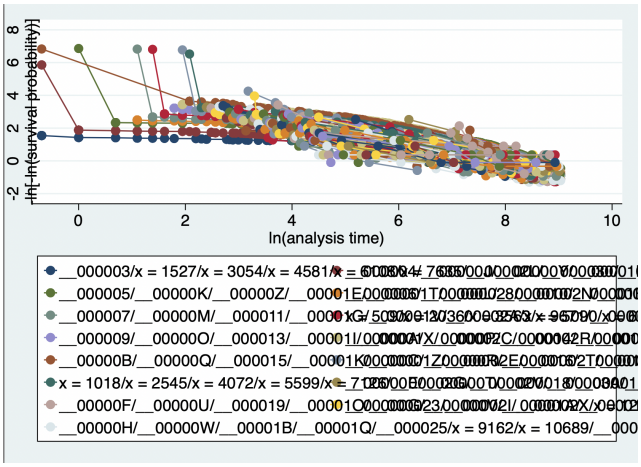
Graph 6: by operation



Graph 7: by length of stay



Graph 8: by cost



Graph 9: by ln(cost)

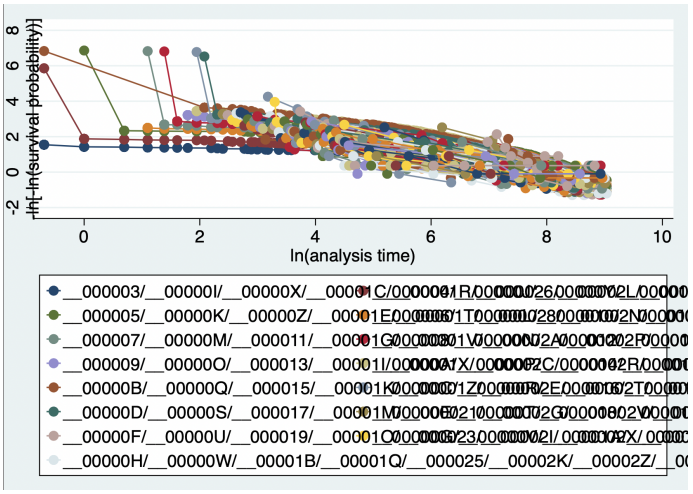


Table 9: Stratified cox proportional hazards model output (stratified by age, using untransformed costs)

```
. stcox female comorb oper los x, strata(ageyrs)

        Failure _d: death==1
  Analysis time _t: time

note: x omitted because of collinearity.
Iteration 0:   log likelihood = -3763.1433
Iteration 1:   log likelihood = -3631.9963
Iteration 2:   log likelihood = -3631.5085
Iteration 3:   log likelihood = -3631.5084
Refining estimates:
Iteration 0:   log likelihood = -3631.5084

Stratified Cox regression with Breslow method for ties
Strata variable: ageyrs

No. of subjects =     14,490                     Number of obs = 14,490
No. of failures =     10,765
Time at risk    = 39,756,053

                                                 LR chi2(4)    =  263.27
Log likelihood = -3631.5084                      Prob > chi2   =  0.0000
```

_t	Haz. ratio	Std. err.	z	P>|z|	[95% conf.	interval]
female	.9628059	.0399176	-0.91	0.361	.8876632	1.04431
comorb	2.080636	.1025914	14.86	0.000	1.888971	2.291748
oper	.8454051	.0771994	-1.84	0.066	.7068644	1.011099
los	1.003248	.0008154	3.99	0.000	1.001651	1.004847
x	1	(omitted)				

```
.         *^higher log likleihood (-3000 versus -25000)
```

```
. estat phtest, detail

Test of proportional-hazards assumption

Time function: Analysis time
```

	rho	chi2	df	Prob>chi2
female	-0.00586	0.38	1	0.5401
comorb	-0.05787	36.21	1	0.0000
oper	0.01094	1.31	1	0.2523
los	0.01310	2.22	1	0.1359
o.x	.	.	1	.
Global test		39.45	4	0.0000

Table 10: Stratified cox proportional hazards model output (stratified by age, using log transformed costs)

```
.  stcox female comorb oper los lnx, strata(ageyrs)

        Failure _d: death==1
  Analysis time _t: time

Iteration 0:   log likelihood = -3763.1433
Iteration 1:   log likelihood = -3631.9728
Iteration 2:   log likelihood = -3631.4868
Iteration 3:   log likelihood = -3631.4867
Refining estimates:
Iteration 0:   log likelihood = -3631.4867

Stratified Cox regression with Breslow method for ties
Strata variable: ageyrs

No. of subjects =     14,490                        Number of obs = 14,490
No. of failures =     10,765
Time at risk    = 39,756,053
                                                    LR chi2(5)    = 263.31
Log likelihood = -3631.4867                         Prob > chi2   = 0.0000
```

_t	Haz. ratio	Std. err.	z	P>|z|	[95% conf.	interval]
female	.9626742	.0399178	-0.92	0.359	.8875317	1.044179
comorb	2.079975	.1026059	14.85	0.000	1.888288	2.291122
oper	.8462522	.0773884	-1.83	0.068	.7073898	1.012374
los	1.003108	.001054	2.95	0.003	1.001044	1.005176
lnx	1.004968	.0239153	0.21	0.835	.9591716	1.052952

```
. estat phtest, detail

Test of proportional-hazards assumption

Time function: Analysis time
```

	rho	chi2	df	Prob>chi2
female	-0.00639	0.45	1	0.5036
comorb	-0.05889	37.50	1	0.0000
oper	0.01273	1.77	1	0.1828
los	-0.00932	1.15	1	0.2834
lnx	0.03210	12.15	1	0.0005
Global test		51.62	5	0.0000

```
.
```

Table 11: Cox proportional hazards model output with time variant costs

```
.
. // time variant cox
.
. stcox ageyrs female comorb oper los, tvc(cost)

        Failure _d: death==1
  Analysis time _t: time

Iteration 0:   log likelihood = -95735.671
Iteration 1:   log likelihood = -93585.936
Iteration 2:   log likelihood = -93545.668
Iteration 3:   log likelihood = -93544.317
Iteration 4:   log likelihood = -93544.313
Refining estimates:
Iteration 0:   log likelihood = -93544.313

Cox regression with Breslow method for ties

No. of subjects =     14,490                     Number of obs =  14,490
No. of failures =     10,765
Time at risk    = 39,756,053
                                                 LR chi2(6)    = 4382.71
Log likelihood = -93544.313                      Prob > chi2   =  0.0000
```

_t	Haz. ratio	Std. err.	z	P>|z|	[95% conf.	interval]
main						
ageyrs	1.055669	.0010396	55.01	0.000	1.053633	1.057708
female	.9397658	.0185397	-3.15	0.002	.9041222	.9768147
comorb	1.650104	.03756	22.00	0.000	1.578106	1.725387
oper	.9093037	.0391985	-2.21	0.027	.8356321	.9894705
los	1.002991	.0004087	7.33	0.000	1.002191	1.003793
tvc						
costofstay	1	2.00e-09	0.86	0.388	1	1

```
Note: Variables in tvc equation interacted with _t.
```

Table 12: Exponential model output with untransformed costs

```
.
. // exponential distribution with untransformed costs
. streg ageyrs female comorb oper los x, distribution(exponential)

        Failure _d: death==1
  Analysis time _t: time
note: x omitted because of collinearity.

Iteration 0:   log likelihood = -27546.114
Iteration 1:   log likelihood = -25601.337
Iteration 2:   log likelihood = -25295.808
Iteration 3:   log likelihood =  -25289.41
Iteration 4:   log likelihood = -25289.407
Iteration 5:   log likelihood = -25289.407

Exponential PH regression

No. of subjects =     14,490                          Number of obs =  14,490
No. of failures =     10,765
Time at risk    = 39,756,053
                                                      LR chi2(5)    = 4513.41
Log likelihood = -25289.407                           Prob > chi2   =  0.0000
```

_t	Haz. ratio	Std. err.	z	P>|z|	[95% conf.	interval]
ageyrs	1.054128	.0010068	55.19	0.000	1.052157	1.056103
female	.9450169	.01863	-2.87	0.004	.9091992	.9822457
comorb	1.641523	.0371103	21.92	0.000	1.570377	1.715894
oper	.8746462	.0376455	-3.11	0.002	.8038887	.9516316
los	1.00341	.0003056	11.18	0.000	1.002811	1.00401
x	1	(omitted)				
_cons	6.14e-06	3.99e-07	-184.60	0.000	5.41e-06	6.97e-06

Note: **_cons** estimates baseline hazard.

Table 13: AIC for exponential model output with untransformed costs

```
.
. estat ic

Akaike's information criterion and Bayesian information criterion
```

Model	N	ll(null)	ll(model)	df	AIC	BIC
.	14,490	-27546.11	-25289.41	6	50590.81	50636.3

Note: BIC uses N = number of observations. See **[R] BIC note**.

Table 14: Exponential model output with log transformed costs

```
. // exponential distribution with log transformed costs
.
. streg ageyrs female comorb oper los lnx, distribution(exponential)

        Failure _d: death==1
  Analysis time _t: time

Iteration 0:   log likelihood = -27546.114
Iteration 1:   log likelihood = -25591.454
Iteration 2:   log likelihood = -25292.195
Iteration 3:   log likelihood = -25285.799
Iteration 4:   log likelihood =  -25285.79
Iteration 5:   log likelihood =  -25285.79

Exponential PH regression

No. of subjects =     14,490                     Number of obs =  14,490
No. of failures =     10,765
Time at risk    = 39,756,053
                                                 LR chi2(6)    = 4520.65
Log likelihood = -25285.79                       Prob > chi2   =  0.0000
```

_t	Haz. ratio	Std. err.	z	P>|z|	[95% conf.	interval]
ageyrs	1.053748	.0010163	54.28	0.000	1.051758	1.055742
female	.9436137	.0186083	-2.94	0.003	.907838	.9807992
comorb	1.634709	.0370458	21.69	0.000	1.563689	1.708954
oper	.8802938	.0379442	-2.96	0.003	.8089794	.9578948
los	1.002715	.0004118	6.60	0.000	1.001908	1.003522
lnx	1.031421	.011941	2.67	0.008	1.008281	1.055093
_cons	5.20e-06	4.68e-07	-135.11	0.000	4.36e-06	6.20e-06

Note: _cons estimates baseline hazard.

Table 15: AIC for exponential model output with log transformed costs

```
. estat ic

Akaike's information criterion and Bayesian information criterion
```

Model	N	ll(null)	ll(model)	df	AIC	BIC
.	14,490	-27546.11	-25285.79	7	50585.58	50638.65

Note: BIC uses N = number of observations. See [R] BIC note.

Graph 10: Exponential regression

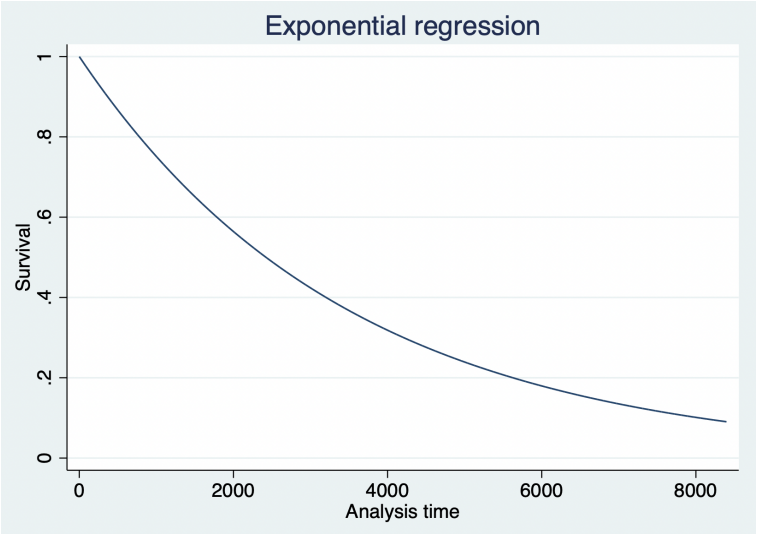


Table 16: Weibull model output with untransformed costs

```
. // Weibull regression with untransformed costs
. streg ageyrs female comorb oper los x, distribution(weibull)

        Failure _d: death==1
  Analysis time _t: time
note: x omitted because of collinearity.

Fitting constant-only model:
Iteration 0:   log likelihood = -27546.114
Iteration 1:   log likelihood =  -26426.54
Iteration 2:   log likelihood = -26399.569
Iteration 3:   log likelihood =  -26399.55
Iteration 4:   log likelihood =  -26399.55

Fitting full model:
Iteration 0:   log likelihood =  -26399.55
Iteration 1:   log likelihood = -24831.953
Iteration 2:   log likelihood = -24675.148
Iteration 3:   log likelihood = -24674.123
Iteration 4:   log likelihood = -24674.121

Weibull PH regression

No. of subjects =     14,490                          Number of obs =  14,490
No. of failures =     10,765
Time at risk    = 39,756,053
                                                      LR chi2(5)    = 3450.86
Log likelihood = -24674.121                           Prob > chi2   =  0.0000
```

_t	Haz. ratio	Std. err.	z	P>|z|	[95% conf.	interval]
ageyrs	1.047809	.0010091	48.49	0.000	1.045833	1.049789
female	.9541071	.0188072	-2.38	0.017	.9179486	.9916898
comorb	1.561432	.0353752	19.67	0.000	1.493615	1.632328
oper	.8498561	.0365826	-3.78	0.000	.7810969	.9246683
los	1.00286	.0003084	9.28	0.000	1.002255	1.003464
x	1	(omitted)				
_cons	.0000697	6.34e-06	-105.26	0.000	.0000584	.0000833
/ln_p	-.2822257	.0085241	-33.11	0.000	-.2989326	-.2655188
p	.7541035	.006428			.7416094	.7668081
1/p	1.326078	.0113036			1.304107	1.348419

Table 17: AIC for weibull model output with untransformed costs

```
. estat ic

Akaike's information criterion and Bayesian information criterion
```

Model	N	ll(null)	ll(model)	df	AIC	BIC
.	14,490	−26399.55	−24674.12	7	49362.24	49415.31

Note: BIC uses N = number of observations. See [R] BIC note.

Table 18: Weibull model output with log transformed costs

```
. // Weibull regression with log transformed costs
. streg ageyrs female comorb oper los lnx, distribution(weibull)

        Failure _d: death==1
  Analysis time _t: time

Fitting constant-only model:
Iteration 0:   log likelihood = -27546.114
Iteration 1:   log likelihood =  -26426.54
Iteration 2:   log likelihood = -26399.569
Iteration 3:   log likelihood =  -26399.55
Iteration 4:   log likelihood =  -26399.55

Fitting full model:
Iteration 0:   log likelihood =  -26399.55
Iteration 1:   log likelihood = -24825.962
Iteration 2:   log likelihood = -24672.069
Iteration 3:   log likelihood = -24671.026
Iteration 4:   log likelihood = -24671.023

Weibull PH regression

No. of subjects =     14,490                     Number of obs =  14,490
No. of failures =     10,765
Time at risk    = 39,756,053
                                                 LR chi2(6)    = 3457.05
Log likelihood = -24671.023                      Prob > chi2   =  0.0000
```

_t	Haz. ratio	Std. err.	z	P>|z|	[95% conf.	interval]
ageyrs	1.047462	.0010183	47.70	0.000	1.045468	1.04946
female	.9527862	.0187873	−2.45	0.014	.9166663	.9903294
comorb	1.555673	.0353224	19.46	0.000	1.48796	1.626467
oper	.8547408	.0368432	−3.64	0.000	.7854957	.9300902
los	1.002213	.0004139	5.35	0.000	1.001402	1.003025
lnx	1.028936	.0118626	2.47	0.013	1.005947	1.052451
_cons	.0000598	6.60e−06	−88.10	0.000	.0000481	.0000742
/ln_p	−.2820308	.0085221	−33.09	0.000	−.2987338	−.2653279
p	.7542504	.0064278			.7417568	.7669545
1/p	1.32582	.0112988			1.303858	1.348151

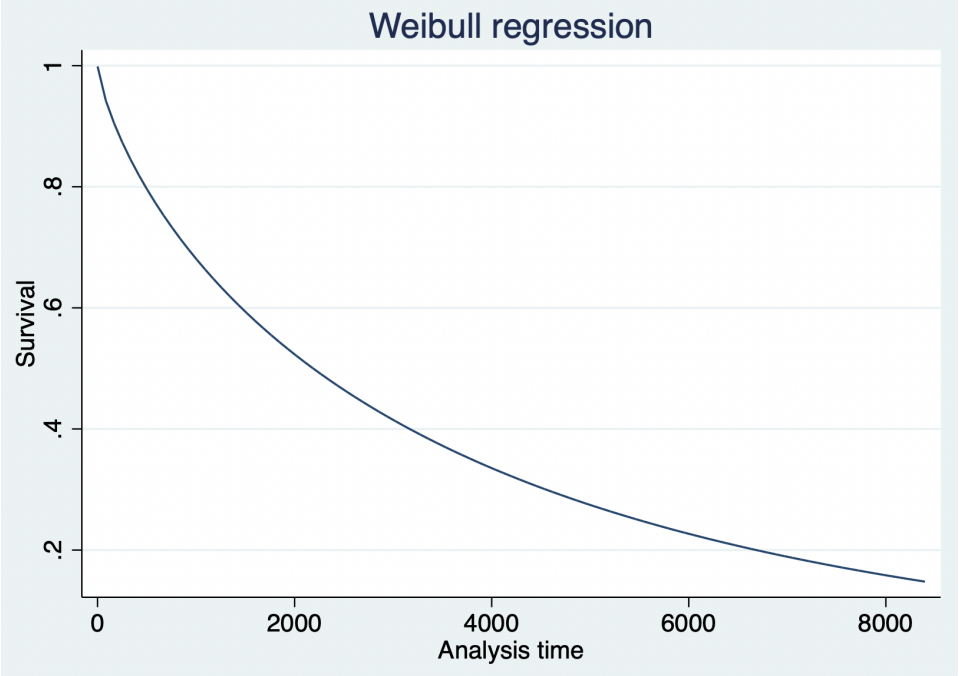
Table 19: AIC for weibull model output with log transformed costs

```
.
. estat ic

Akaike's information criterion and Bayesian information criterion
```

Model	N	ll(null)	ll(model)	df	AIC	BIC
.	14,490	-26399.55	-24671.02	8	49358.05	49418.7

Note: BIC uses N = number of observations. See [R] BIC note.

Graph 11: Weibull regression



Citations

1. Anderson, M. L., Dobson, S. D., & Reinsel, G. C. (2019). Long-term climate forcing by atmospheric oxygen concentrations. Earth and Planetary Science Letters, 507, 17-25. https://doi.org/10.1016/j.epsl.2018.11.006
2. Bosman, P.A.N., and Thierens, D. "Negative log-likelihood and statistical hypothesis testing as the basis of model selection in IDEAs." Utrecht University Repository, 2000. (Preprint). URL: https://dspace.library.uu.nl/handle/1874/1969
3. Chughtai M, Gwam CU, Mohamed N, et al. (2018). "Surgical site infection and hospital readmission rates after total hip arthroplasty: a six-year analysis." J Bone Joint Surg Am, 100(22):1913-1920. doi: 10.2106/JBJS.17.01613. PMID: 30451729.
4. Crimmins, Eileen M., & Goldman, Noreen. The Male-Female Health-Survival Paradox: A Comparative Perspective on Sex Differences in Aging and Mortality. Bethesda (MD): National Center for Biotechnology Information (US), 2016. https://www.ncbi.nlm.nih.gov/books/NBK242444/.

5. de Castro, Samuel Mendes, et al. "Surgical treatment of gangrenous cholecystitis: evaluation of the Tokyo Guidelines 2018." BMC Surgery, 22, 67 (2022). https://doi.org/10.1186/s12893-022-01835-1.
6. Gagnon, Alain, & Mazan, Ryan. "Women live longer than men even during severe famines and epidemics." Proceedings of the National Academy of Sciences, 114(37), E7892-E7894 (2017). https://doi.org/10.1073/pnas.1701535115.
7. Kim, Jiyun, et al. "Exploring the relationship between integrated care and the experiences of people living with multimorbidity: a systematic review." BMC Health Services Research, 18, 570 (2018). https://doi.org/10.1186/s12913-018-2916-1.
8. Kravdal, Ø., & Kravdal, H. (Eds.). The Male-Female Health-Survival Paradox: A Comparative Perspective on Sex Differences in Aging and Mortality. Bethesda (MD): National Center for Biotechnology Information (US), 2016. https://www.ncbi.nlm.nih.gov/books/NBK242444/.
9. Lindsey, J. K., & Jones, B. "Choosing among generalized linear models applied to medical data." Statistics in Medicine, 17(1), 1998, 59-68. DOI: https://doi.org/10.2307/1911033
10. Marsh, David J., et al. "Prognostic factors in small cell lung cancer: an analysis of 1,521 patients." American Journal of Medicine, 93(6), 615-622 (1992). https://doi.org/10.1016/0002-9343(92)90166-a.
11. Meara JG, Leather AJ, Hagander L, et al. (2015). "Global surgery 2030: evidence and solutions for achieving health, welfare, and economic development." Lancet, 386(9993):569-624. doi:

Stata Code

```
use "/Users/sophialind/Downloads/HP425_summative dataset_2022-23.dta", clear

///Formatting in Stata's date format.
format mineadmdate %td
format edeathdate %td
format elastdate %td

describe

///Creating a variable called "time" that gives the time from start of study to failure/death or censoring
///(whichever comes first). --time in the study (die or finish)
cap drop time
gen time = .
replace time = edeathdate - mineadmdate if edeathdate != .
replace time = elastdate - mineadmdate if edeathdate == .

* Replacing any observed zero values for "time" with a value of 0.5.
replace time = 0.5 if time == 0

clonevar y=time
summarize y, detail
```

```
table time

///Create a clone variable (x) for cost
clonevar x=costofstay
summarize x, detail
table costofstay
*data is skewed right

///Check if there are missing values (no missings found in the data)
drop if x==0

///Graphical representation of the data
twoway histogram x, color(*0.5)||kdensity x , title(costofstay data)

///obtaining detailed summary of independent variables:
summarize los ageyrs female comorb oper costofstay agedeathyrs

///Transforming the costperday data to adjust for skewness
//Create new variable sqrty for sqrt transformation and and logy for log trnasformation of
costperday data
generate sqrtx=x^0.5
generate lnx=ln(x)

//Graphical representation of the transformed data
twoway histogram sqrty, color(*0.5)||kdensity sqrty, title(cost-per-day data: Square root
transformation)
twoway histogram lny, color(*0.5)||kdensity lny, title(cost-per-day data: Log transformation)

//Create normal probability plots of costperday
pnorm y,title(normal plot of costs of stay) ytitle(costs) xtitle(inverse normal)
pnorm sqrty,title(normal plot of square root costs of stay) ytitle( sqrt costs) xtitle(inverse
normal)
pnorm lny,title(normal plot of log costs of stay) ytitle( log costs) xtitle(inverse normal)

//Obtain detail summary of sqrty and logy for comparison
summarize sqrtx,detail
summarize lnx,detail


///Investigating relationship between mean and variance in the raw data to check for
heteroskedasticity and represent it graphically

//level of cost
global xvars "los ageyrs female comorb oper x agedeathyrs"
global qy=20
quietly regress y $xvars
predict yf_q,xb
```

```
xtile yq=yf_q, nq($qy)
generate yqmean=0
generate yqvar=0
forvalues i=1/$qy {
quietly summarize y if yq==`i',detail
replace yqmean=r(mean) if yq==`i'
replace yqvar=r(Var) if yq==`i'
}
twoway scatter yqvar yqmean,title(Mean and Variance) ytitle(variance) xtitle(mean)

//log of costs

global xvars "los ageyrs female comorb oper lnx agedeathyrs"
global qy1=20
quietly regress y $xvars
predict yf_q1,xb

xtile yq1=yf_q1, nq($qy1)
generate yqmean1=0
generate yqvar1=0
forvalues i=1/$qy {
quietly summarize y if yq1==`i',detail
replace yqmean1=r(mean) if yq1==`i'
replace yqvar1=r(Var) if yq1==`i'
}
twoway scatter yqvar1 yqmean1,title(Mean and Variance) ytitle(variance) xtitle(mean)

///Check the data for censoring
///prepare the data for survival analysis
stset time, failure(death==1)
tab death
*death not observed in 3,725 pts

///Estimate means and associated confidence intervals of survival time
stci, rmean
stci, emean
sts

///See if the time is censored or skewed
sum time
sum time if death==1
sum time if death==0
hist time


///See if the costofstay is censored or skewed
sum costofstay
sum costofstay if death==1
```

```
sum costofstay if death==0
hist costofstay

///OLS
/// OLS regression with untransformed costs
global xvars "los ageyrs female comorb oper x agedeathyrs"

regress y $xvars, robust
predict yhat, xb
summarize yhat, detail
correlate y yhat
spearman y yhat
//model diagnostic test to detect heteroskedasticity
quietly regress y $xvars
estat hettest
// model diagnostic to detect reliability
estat ovtest

//test null hypothesis that coefficient of square of fitted values is not significantly different to
zero
linktest


/// OLS regression with log transformation of costs
global xvars "los ageyrs female comorb oper lnx agedeathyrs"
regress y $xvars, robust
predict yhat1, xb
summarize yhat1, detail
correlate y yhat1
spearman y yhat1
//model diagnostic test to detect heteroskedasticity
quietly regress y $xvars
estat hettest
// model diagnostic to detect reliability
estat ovtest

//test null hypothesis that coefficient of square of fitted values is not significantly different to
zero
linktest

// cox proportional hazards model with untransformed costs
stset time, failure(death==1)

stcox ageyrs female comorb oper los x
stcox, nohr

* Graph the resulting output:
stcurve, survival title(Cox proportional hazards model (untransformed costs))
```

```
// cox proportional hazards model with log transformation of costs

stcox ageyrs female comorb oper los lnx
stcox, nohr
stcurve, survival title(Cox proportional hazards model (log transformed costs))

/// Test for equality of survival functions between patients in different groups (Proportional
hazard assumption)
sts test comorb, logrank
sts graph, by(comorb)
graph export "${home}/km_survival_curve.png", replace

///oper
sts test oper, logrank
sts graph, by(oper)
graph export "${home}/km_survival_curve.png", replace

///female
sts test female, logrank
sts graph, by(female)
raph export "${home}/km_survival_curve.png", replace

///ageyrs
sts test ageyrs, logrank
sts graph, by(ageyrs)
graph export "${home}/km_survival_curve.png", replace

///cost
sts test x, logrank
sts graph, by(x)
graph export "${home}/km_survival_curve.png", replace

///TESTING Porportional hazard assumption by each covariate
stphplot, by(female)
stphplot, by(comorb)
stphplot, by(oper)
stphplot, by(los)
stphplot, by(x)
stphplot, by(lnx)

stphplot, by(ageyrs)
        *^error; no room to add more variables

estat phtest
estat phtest, detail

// cox proportional hazards model with stratification
```

```
stcox ageyrs comorb oper los x, strata(female)

stcox ageyrs female comorb oper los x, strata(edeathdate)
estat phtest, detail
///prop assum holds except for comorb, los

stcox ageyrs female comorb oper x, strata(comorb)
estat phtest, detail

stcox female comorb oper los x, strata(ageyrs)
        *^higher log likleihood (-3000 versus -25000)
        **prop assum holds except for comorb, los
estat phtest, detail


stcox female comorb oper lnx, strata(ageyrs)
        *^higher log likleihood (-3000 versus -25000)
estat phtest, detail

stcox female comorb oper x, strata(ageyrs)
        *^higher log likleihood (-3000 versus -25000)
estat phtest, detail


stcox female comorb oper los, strata(ageyrs) tvc(cost)
estat phtest, detail


// Create a scatter plot of "cost" over time
twoway scatter cost time

// time variant cox

stcox ageyrs female comorb oper los, tvc(cost)
        *log likelihood -900000 (very low --> suggests not a good fit)
stcox ageyrs female oper los, tvc(comorb)

// exponential distribution with untransformed costs
streg ageyrs female comorb oper los x, distribution(exponential)
streg, nohr
estat ic

// exponential distribution with log transformed costs
streg ageyrs female comorb oper los lnx, distribution(exponential)
streg, nohr
estat ic
```

```
/// Save the predicted mean survival time. Obtain the mean of predicted
/// mean survival time by summarising this new variable.
predict survmean_exp, mean time
summ survmean_exp

graph twoway ///
(scatter survmean_exp ageyrs if female==1, msize(vtiny) mcolor(red)) ///
(scatter survmean_exp ageyrs if female==0, msize(vtiny) mcolor(blue))
graph export "${home}/3_survmean_exp.png", replace

/// Graph the resulting output :
stcurve, survival
graph export "${home}/3_exponential.png", replace

// Weibull regression with untransformed costs
streg ageyrs female comorb oper los x, distribution(weibull)
streg, nohr
estat ic

// Weibull regression with log transformed costs
streg ageyrs female comorb oper los lnx, distribution(weibull)
streg, nohr
estat ic

/// Save the predicted mean survival time from the Weibull regression to a new
/// variable survmean_wei:
predict survmean_wei, mean time
summ survmean_wei

/// Graph the resulting output using the following commands:
stcurve, survival
graph export "${home}/4_weibull.png", replace
```