University of California, Santa Barbara
PSTAT 175 Final Project

Do Social and Economic Status Affect Mortality Age?

Inwoong Bae, Sarah Hermann, Ghazaleh Moradi
Group 1
Instructor: Adam Tashman
December 6, 2019

INTRO

In this project, we build an appropriate survival model to determine which factors (educational level, electricity, sex, and marriage) affect the age of death of people in Bihar, a state in India, in 2007.

DATASET

Our data is from Kaggle.com about mortality schedules and other factors from the Empowered Action Group (EAG) states in India. We decided to specify our data to one specific state, Bihar and deaths that occurred in one year, 2007 because the whole data set contained 770,000 observations from 9 different states over about 6 years. The data set had 121 variables so we narrowed our data set down to 9 variables (ID, Sex, Age of Death, marital status, highest Education, Electricity, and our added variables that we filtered: Censored Age of Death, Marital Status Category and Highest Education Category). We altered multiple of the variables by the following:

- **Marital Status:** Marital Status began with 7 different levels, but many had the same or similar outcome; for example, there was a category for married with a gauna as well as married without a gauna. We focused on whether or not the subject was married versus not as the smaller details were less important in our analysis. We assigned the values 1 to any person who was never married, 2 to any person who was married without a gauna performed or married with a gauna performed or remarried, and 3 to any person who was a widow/widower or divorced or separated.
- **Education**: The variable initially had 9 different levels of education but we decided that having a college education versus not was a more important split. We assigned the value 1 to any person with education less than college/university (ranges from illiterate to

high school) and 2 to anyone with a college education or higher. NA/NULL values have been omitted.

- **Censored Age of Death**: If the subject was under 24 years old, we censored their death because they were often not married, all had very similar education levels and often may have died really young and were not affected by any of these factors.
- **Households have electricity**: Households with electricity was already formatted with 1 equaling having electricity and 2 meaning they do not have electricity.

Our other variables are:

- Sex: Male=1, Female=2
- Age of Death
- Marital Status--which we manipulated into Marital Status Category
- Highest Education--which was manipulated into Highest Education Category

We also omitted any NA/NULL values to avoid missing values and also because we would not be able to give accurate estimates for them.

	ID <int>	Sex <int>	Age of Death <int>	martial Status <int>	Highest Education <int>	Household have electricity <int>	Censored Age of Death <dbl>	martial Status category <dbl>	Highest Education Category <dbl>
1	3264	2	82	3	1	2	1	2	1
2	3272	1	48	5	2	2	1	3	1
3	3336	2	84	3	2	2	1	2	1
4	3384	1	62	3	1	2	1	2	1
5	3392	1	50	5	1	1	1	3	1
6	3120	1	24	3	1	2	0	2	1

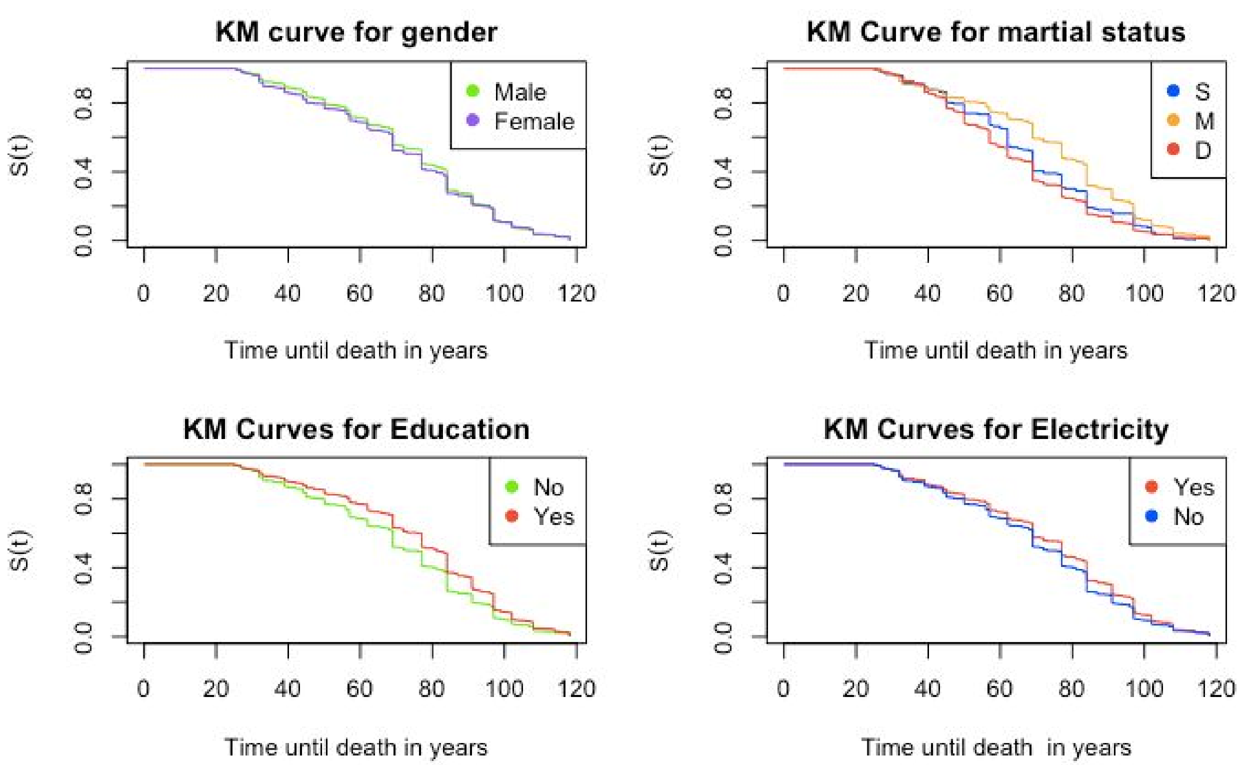
6 rows

QUESTION

We are interested in whether gender, marital status, level of education, and/or access to electricity or any combination of these factors had an influence on the age of death in that specific state. In addition, we want to see if there is any interaction between any of these covariates: sex, martial status, highest education, and electricity.

KAPLAN MEIER

We ran a Kaplan Meier test for age of death against sex, marital status and electricity. We found that males tend to have a slightly higher survival rate than women. Subjects that had electricity were likely to survive longer than subjects without electricity. Subjects who were married had a much higher rate of survival than subjects who were single, and subjects who were divorced, widowed, or separated had a much lower survival rate than most. The biggest difference can be seen in the marital status plot, so the different levels of marriage lead to the largest difference in age of death between the categories. Below the Kaplan Meier curves, the p-values for each of the survdiff tests can be seen.



```
Call:
survdiff(formula = Surv(time_BH, status_BH) ~ sex_BH)

             N Observed Expected (O-E)^2/E (O-E)^2/V
sex_BH=1 9329     8072     8192      1.76      4.72
sex_BH=2 7844     6586     6466      2.22      4.72

 Chisq= 4.7  on 1 degrees of freedom, p= 0.03
Call:
survdiff(formula = Surv(time_BH, status_BH) ~ martial_BH)

                 N Observed Expected (O-E)^2/E (O-E)^2/V
martial_BH=1   157      146      117      7.32      8.45
martial_BH=2 13869    11556    12481     68.56    538.82
martial_BH=3  3147     2956     2060    389.49    528.72

 Chisq= 543  on 2 degrees of freedom, p= <2e-16
Call:
survdiff(formula = Surv(time_BH, status_BH) ~ edu_BH)

              N Observed Expected (O-E)^2/E (O-E)^2/V
edu_BH=1 13903    11749    11264      20.9       108
edu_BH=2  3270     2909     3394      69.3       108

 Chisq= 108  on 1 degrees of freedom, p= <2e-16
Call:
survdiff(formula = Surv(time_BH, status_BH) ~ elec_BH)

               N Observed Expected (O-E)^2/E (O-E)^2/V
elec_BH=1  6348     5499     5903      27.7      55.1
elec_BH=2 10825     9159     8755      18.6      55.1

 Chisq= 55.1  on 1 degrees of freedom, p= 1e-13
```

MODEL BUILDING

Now, we need to build our Cox PH model. We use both backward elimination as well as forward stepwise to find out the best covariates for our model. Our full model consists of four covariates (gender, education level, marital status and electricity). In backward elimination method, the function stops with the model with all four variables. We repeat this step using a different method (forward stepwise function) to be sure we have the best possible model. Since we approach the result again, we decide to have a full model with all four covariates.

```
step(Full_model, direction = "backward")
```

```
Start:  AIC=251428
Surv(time_BH, status_BH) ~ sex_BH + edu_BH + elec_BH + martial_BH

             Df    AIC
<none>          251428
- sex_BH      1 251434
- elec_BH     1 251462
- edu_BH      1 251468
- martial_BH  1 251785
```

Second, we try our full model in both likelihood tests and anova table to be sure the covariates will pass the tests and we do not include unnecessary predictor to our model.

LIKELIHOOD TESTS

H_0: the reduced model is preferred (with smallest covariates)

H_a: the full model is preferred

Since the resulting p-value in likelihood tests is 0.0048, which is less than 0.05, it indicates that the model with all four predictors fits significantly better than reduced model.

```
pchisq(lrt3,df=1,lower.tail=FALSE)
```
```

 [1] 0.004805534
```

ANOVA TABLE

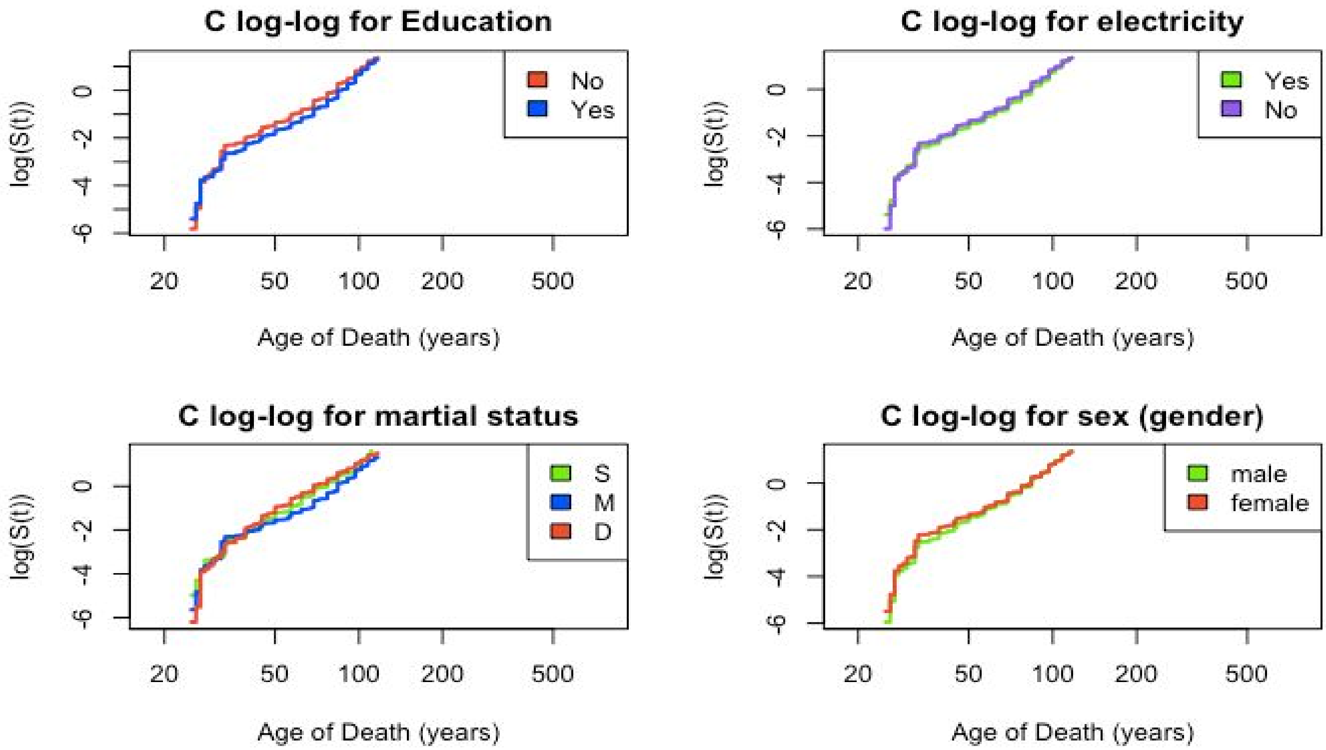
We also tested the full model with anova and the result shows all predictors except sex are significant but becasue the sex predictor was not rejected by any other tests, we decided to keep it for now and do more tests in our full model.

:

```
             loglik    Chisq Df Pr(>|Chi|)    
NULL        -125958                           
sex_BH      -125956   3.6869  1    0.05484 .  
edu_BH      -125904 104.0815  1  < 2.2e-16 ***
elec_BH     -125890  29.1045  1  6.858e-08 ***
martial_BH  -125710 359.3735  1  < 2.2e-16 ***
```

C-LOG-LOG PLOTS

From our C-log-log plots, we notice that all 4 curves are parallel throughout most of the graph with the exception of a small portion at the beginning, so we can conclude, based on the plots as well as the p-values from the cox.zph(), that the assumption is not satisfied because the cox.zph() p-values are less than 0.05.




```
             chisq df       p
edu_BH        33.9  1 5.9e-09
martial_BH    63.4  1 1.7e-15
elec_BH       10.3  1  0.0013
sex_BH        28.9  1 7.5e-08
```

Since all the C-log-log plots resulted in p-values that were less than 0.05, we had to look into stratifying some variables or creating interaction terms. We first looked at stratifying the variable with the lowest p-values, marital status, in hopes that it was the one that was messing up the PH assumption. Instead, we tried building another model using a stratified model with an interaction term.

TESTING FOR STRATIFICATION AND INTERACTION TERMS

```
                                       rho  chisq        p
edu_BH                             0.00823  1.039 3.08e-01
elec_BH                           -0.01361  2.849 9.14e-02
sex_BH                            -0.01000  1.629 2.02e-01
strata(martial_BH)martial_BH=2:edu_BH  -0.00364  0.203 6.53e-01
strata(martial_BH)martial_BH=3:edu_BH  -0.01054  1.695 1.93e-01
strata(martial_BH)martial_BH=2:elec_BH  0.01077  1.784 1.82e-01
strata(martial_BH)martial_BH=3:elec_BH  0.01390  2.969 8.49e-02
strata(martial_BH)martial_BH=2:sex_BH   0.00543  0.481 4.88e-01
strata(martial_BH)martial_BH=3:sex_BH   0.00745  0.901 3.43e-01
GLOBAL                                  NA 83.513 3.23e-14
Call:
coxph(formula = Surv(time_BH, status_BH) ~ strata(martial_BH) *
    edu_BH + strata(martial_BH) * elec_BH + strata(martial_BH) *
    sex_BH)

  n= 17173, number of events= 14658
```

```
                                             coef exp(coef)  se(coef)      z Pr(>|z|)
edu_BH                                 -7.738e-03 9.923e-01 1.941e-01 -0.040   0.9682
elec_BH                                 3.519e-01 1.422e+00 1.829e-01  1.925   0.0543 .
sex_BH                                  7.053e-02 1.073e+00 1.821e-01  0.387   0.6986
strata(martial_BH)martial_BH=2:edu_BH  -1.417e-01 8.679e-01 1.954e-01 -0.725   0.4685
strata(martial_BH)martial_BH=3:edu_BH  -1.712e-05 1.000e+00 2.041e-01  0.000   0.9999
strata(martial_BH)martial_BH=2:elec_BH -2.428e-01 7.844e-01 1.839e-01 -1.320   0.1868
strata(martial_BH)martial_BH=3:elec_BH -2.745e-01 7.599e-01 1.868e-01 -1.469   0.1418
strata(martial_BH)martial_BH=2:sex_BH  -1.548e-02 9.846e-01 1.831e-01 -0.085   0.9326
strata(martial_BH)martial_BH=3:sex_BH  -4.054e-02 9.603e-01 1.860e-01 -0.218   0.8275
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                                       exp(coef) exp(-coef) lower .95 upper .95
edu_BH                                    0.9923     1.0078    0.6784     1.452
elec_BH                                   1.4218     0.7033    0.9936     2.035
sex_BH                                    1.0731     0.9319    0.7509     1.533
strata(martial_BH)martial_BH=2:edu_BH     0.8679     1.1522    0.5918     1.273
strata(martial_BH)martial_BH=3:edu_BH     1.0000     1.0000    0.6703     1.492
strata(martial_BH)martial_BH=2:elec_BH    0.7844     1.2748    0.5470     1.125
strata(martial_BH)martial_BH=3:elec_BH    0.7599     1.3159    0.5269     1.096
strata(martial_BH)martial_BH=2:sex_BH     0.9846     1.0156    0.6878     1.410
strata(martial_BH)martial_BH=3:sex_BH     0.9603     1.0414    0.6668     1.383

Concordance= 0.539  (se = 0.003 )
Likelihood ratio test= 111.8  on 9 df,   p=<2e-16
Wald test            = 109.3  on 9 df,   p=<2e-16
Score (logrank) test = 109.6  on 9 df,   p=<2e-16
```

With each variable, we can see all variates are more than 0.05 and it means that all variates are valid for PH assumption.

However, when we run the summary for the stratified model with interactions, all of them are not significant in this model since P-values of them are over 0.05.

To substitute our model, we built the accelerated failure time model using the weibull distribution, the log-logistic distribution and the exponential distribution.

```
Call:
survreg(formula = Surv(time_BH, status_BH) ~ edu_BH + martial_BH +
    elec_BH + sex_BH, dist = "weibull")
               Value Std. Error       z       p
(Intercept)  4.65937    0.01939  240.30 < 2e-16
edu_BH       0.03532    0.00586    6.03 1.6e-09
martial_BH  -0.11365    0.00566  -20.07 < 2e-16
elec_BH     -0.02721    0.00478   -5.69 1.3e-08
sex_BH      -0.01404    0.00457   -3.07  0.0021
Log(scale)  -1.29126    0.00663 -194.86 < 2e-16

Scale= 0.275

Weibull distribution
Loglik(model)= -66476.3   Loglik(intercept only)= -66726.1
        Chisq= 499.55 on 4 degrees of freedom, p= 8.4e-107
Number of Newton-Raphson Iterations: 12
n= 17173
```

In this model with the weibull distribution, all the variates are significant because the p-values of the variables are less than 0.05. The p-value for the likelihood ratio test is also less than the significant level 0.05. This means this model is statistically significant.

```
Call:
survreg(formula = Surv(time_BH, status_BH) ~ edu_BH + martial_BH +
    elec_BH + sex_BH, dist = "loglogistic")
               Value Std. Error       z       p
(Intercept)  4.63685    0.02454  188.96 < 2e-16
edu_BH       0.05402    0.00743    7.27 3.5e-13
martial_BH  -0.15234    0.00715  -21.32 < 2e-16
elec_BH     -0.03514    0.00615   -5.71 1.1e-08
sex_BH      -0.03197    0.00589   -5.43 5.7e-08
Log(scale)  -1.58785    0.00694 -228.83 < 2e-16

Scale= 0.204

Log logistic distribution
Loglik(model)= -67957.8   Loglik(intercept only)= -68267.8
        Chisq= 619.92 on 4 degrees of freedom, p= 7.6e-133
Number of Newton-Raphson Iterations: 6
n= 17173
```

```
Call:
survreg(formula = Surv(time_BH, status_BH) ~ edu_BH + martial_BH +
    elec_BH + sex_BH, dist = "exponential")
              Value Std. Error     z       p
(Intercept)  4.6668     0.0697 66.91 < 2e-16
edu_BH       0.0376     0.0213  1.76   0.078
martial_BH  -0.1464     0.0202 -7.24 4.5e-13
elec_BH     -0.0284     0.0174 -1.63   0.104
sex_BH      -0.0226     0.0166 -1.36   0.175

Scale fixed at 1

Exponential distribution
Loglik(model)= -77867.5   Loglik(intercept only)= -77899.3
      Chisq= 63.55 on 4 degrees of freedom, p= 5.2e-13
Number of Newton-Raphson Iterations: 4
n= 17173
```

In this model with log-logistic distribution, all the variates are also significant since the p-values of the variables are less than 0.05. The p-value for the likelihood ratio test is also less than the significant level 0.05. This means this model is statistically significant.

In this model with the log-logistic distribution, only martial_BH is significant since the p-value of martial_BH is less than 0.05, but others are greater than the significant level. The p-value for the likelihood ratio test is also less than the significant level 0.05. This means this model is statistically significant.

Based on the above tests, we conclude that the Weibull model is the best choice and our final model.

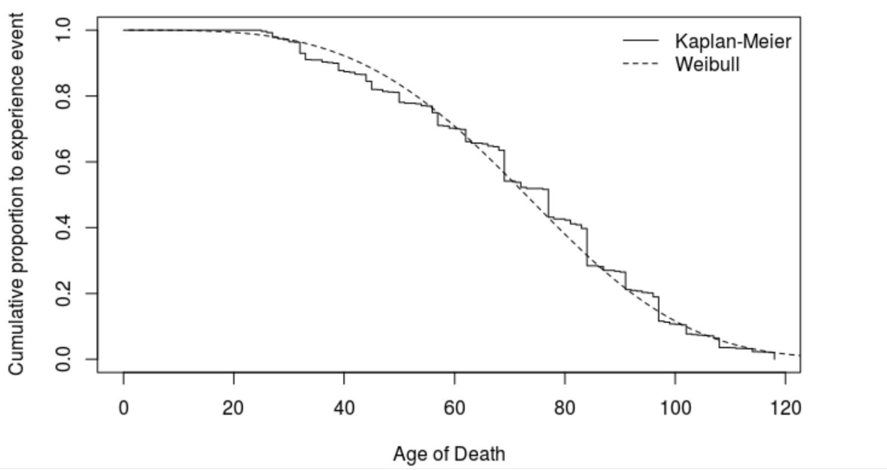
CONFIDENCE INTERVAL FOR HAZARD RATIO

From the AFT model using weibull distribution, the confidence interval for education is in between 0.843470 and 0.9169517. The confidence interval for martial status is in between 1.451878 and 1.5745008. The confidence interval for electricity lies in between 1.067021 and 1.1423352. The confidence interval for sex shows in between 1.018629 and 1.0872495. All of intervals stand for 95% confidence level.

```
$HR
                  HR       LB        UB
edu_BH     0.8794437 0.843470 0.9169517
martial_BH 1.5119467 1.451878 1.5745008
elec_BH    1.1040360 1.067021 1.1423352
sex_BH     1.0523800 1.018629 1.0872495
```

EXTENSION--PARAMETRIC SURVIVAL MODEL

We compare the Kaplan-Meier (non-parametric curve) with Weibull estimates (parametric curve). Since our data did not fit the Kaplan Meier well, we used the Weibull distribution and can see in the below graph that our Weibull model is a good fit for the Kaplan Meier curve. Therefore, it confirms that a Weibull model allows our data to fit.



We used the parametric survival model (psm) function from the RSM package to fit the AFT model and get a closer look at the difference in age of death between sexes.

We looked at the difference in age of death between males and females because on the Kaplan Meier plot, the lines are very close together so we wanted to try to get a better approximation of the difference between sexes. After running the test, we determine that the difference between male and female ages of death is about 4.65 years.

```
addict.aft01 <- psm(Surv(time_BH, status_BH) ~ sex_BH+edu_BH+elec_BH+martial_BH, dist ="weibull")
addict.aft01$coefficients[2]
log.t <- as.numeric(addict.aft01$coefficients[1] + (addict.aft01$coefficients[2]* 1))
log.t
```

```
    sex_BH
-0.0140361
[1] 4.645333
```

EXTENSION--GLM FIT

A generalized linear model (GLM) is used when your response variable does not fit a normal model (which ours does not per all the tests that we used above to try to fit it to a model). When using Age of Death as the response variable, the GLM model with the same covariates that we determined were significant from the Weibull model are also significant. We attempted other GLM models with different combinations of the covariates but it gave us different outputs and determined that many covariates were not significant. The GLM model shows that there is another way (besides all the tests that we learned in class) to come to the same conclusion.

When our 4 covariates all failed the coxph test earlier in the project, we had to turn to trying stratifying and interaction terms and different models to find out that a Weibull model worked best, but with the GLM function, it showed the same result we ended at in much shorter time.

Additionally, from the GLM output, we can interpret the coefficients of each of the covariates, which are the following:

```
Call:
glm(formula = BH2007_filter$`Age of Death` ~ martial_BH + edu_BH +
    elec_BH + sex_BH, data = BH2007_filter)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-71.698  -22.737    5.608   22.838   59.952

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  69.5115     1.9464  35.712  < 2e-16 ***
martial_BH   -2.2300     0.5737  -3.887 0.000102 ***
edu_BH        6.6891     0.5998  11.151  < 2e-16 ***
elec_BH      -2.6171     0.4849  -5.398 6.85e-08 ***
sex_BH       -3.1144     0.4605  -6.763 1.39e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 901.3129)

    Null deviance: 15702504  on 17172  degrees of freedom
Residual deviance: 15473741  on 17168  degrees of freedom
AIC: 165584

Number of Fisher Scoring iterations: 2
```

These coefficients tells us whether the covariates have a positive or negative affect on the response variable, Age of Death. For example, education has a coefficient of 6.6891, meaning that for a one unit increase in education (from less than college education (1) to college education or higher (2)) there is almost a 7 year increase in age of death. And for households that have electricity, our numbering is backwards (1=having electricity and 2=not having electricity) so the negative coefficient makes sense because increasing one unit (to not having electricity) will lead to an age of death almost 3 years less than an age of death with electricity.

CONCLUSION

Our project took a very large, complicated data set and simplified the variables into ones that we believed would have an impact age of death of the subjects. We plotted Kaplan Meier plots to see that most of the covariates were very similar while marital status lead to the largest difference between levels. We then moved to testing the backwards approach and determined that the full model was appropriate. Next, we conducted a likelihood ratio test on the full model and determined that it was preferred to a reduced model. When we plotted the C-log-log plots, they showed a few small crosses in the beginning but were otherwise parallel and did not diverge so we were inclined to say that the PH assumption was not violated. However, when we ran the cox.zph(), all of the covariates had p-values that were less than 0.05, which caused us to conclude that they all violated the PH assumption. Therefore, we turned to trying to stratify our covariates or create interaction terms but to no success, so we tried different models. Eventually we concluded that the Weibull model including all 4 covariates was the best fit for the data.

REFERENCES

Ilangovan, R. (2019). *Predict Mortality/Death Rate.*. [online] Kaggle.com. Available at: https://www.kaggle.com/rajanand/mortality#Mortality_05_UT.csv [Accessed 6 Dec. 2019].

Stevenson, M. (2007). *An Introduction to Survival Analysis*. [online] Biecek.pl. Available at: http://www.biecek.pl/statystykaMedyczna/Stevenson_survival_analysis_195.721.pdf [Accessed 5 Dec. 2019].