Survival Analysis of Worsening Hypertension

PSTAT 196 Report

John Randazzo, Jason Freeberg, Ziyi Jiang

4/16/2017

Introduction

The following is a full report on our data analysis project for the course PSTAT 196: Projects in Actuarial Science. Our analysis was completed during the Spring Quarter of 2017 under the supervision of Professor Ian Duncan and graduate student Shannon Nicponski at The University of California, Santa Barbara.

About the Data

Our dataset comes from a longevity study performed in the United Kingdom from March 2006 to May 2014. Our team's subset of the original data is concerned only with high-risk patients developing severe hyper-tension. All participants in our dataset were already diagnosed with moderate hypertension. Our task was to model the time until diagnosis of sever hypertension–which lends itself to a classic Survival Analysis problem. Our dataset contains 126665 observations and 13 variables. A complete breakdown of the the variables is provided below.

Table 1. A breakdown of all covariates in the original dataset

Name	Explanation	Range
UNQID	A unique ID for each individual	
Sex	The patient's sex	1:
Socio-Economic	A categorical range indicating economic status	1: Affluent –> 5: Deprived
Socio-Econdate	The date of the patient's economic evaluation	
EventDate	The patient's event date	
BMI	The Body Mass Index of the individual	
Alcohol	A binary indicator for drinking habits	1: Drinker, 0: Non-Drinker
Cigar	A binary indicator for smoking habits	1: Smoker, 0: Non-Smoker
YOB	The year of birth of the individual	
Duration	Time in days spent in the study	[0, ???]
Partial_code_pre	Enter code	Always 2000
Partial_code_ff	Transition Code	2000 := Censored,
.	.	3000 := Transitioned
Age_Pre	Age upon entry to the study	
Age_Post	Age at the EventDate	

As you can see, some variables (such as YOB, duration, Age_Pre, Age_Post) are redundant. When our team began the modeling phase, we did not use every available covariate listed in the above table.

Objectives

Our team set the following objectives to complete by the end of the quarter.

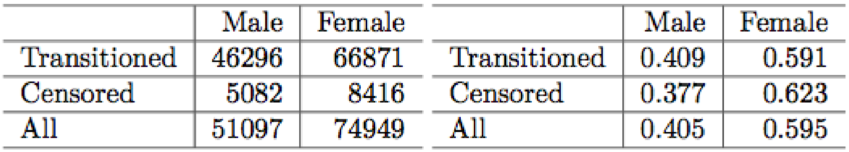
1. Thoroughly explore the dataset for discrepancies or interesting trends
   - Make adjustments as necessary
2. Fit a Cox-PH model to the data
   - Test model assumptions through Schoenfeld Residuals and Log-Log plots
   - Narrow model to the most significant predictors
3. Attempt k-fold cross validation of the decided Cox model
4. If time allows, we chose to pursue the following as well:
   - Non-parametric modeling techniques

Exploratory Analysis

Once our group recieved the dataset, we thuroughly explored the variables using two-way tables and visualizations.

Tabluar Analysis

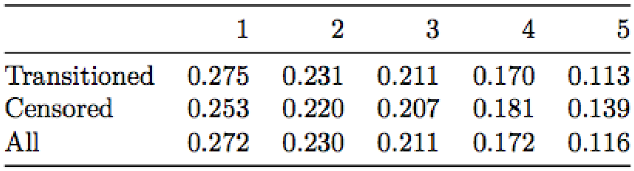
Table 2. Censorship by Sex



	Male	Female
Transitioned	46296	66871
Censored	5082	8416
All	51097	74949

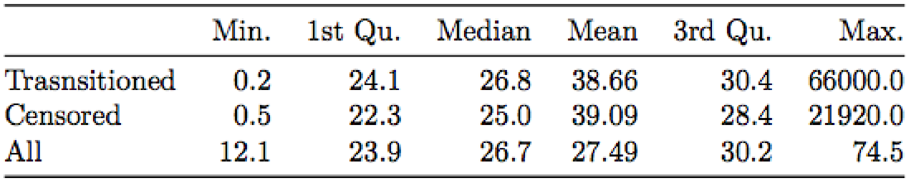
	Male	Female
Transitioned	0.409	0.591
Censored	0.377	0.623
All	0.405	0.595

Table 3. Censorship by Socioeconomic Status



	1	2	3	4	5
Transitioned	0.275	0.231	0.211	0.170	0.113
Censored	0.253	0.220	0.207	0.181	0.139
All	0.272	0.230	0.211	0.172	0.116

Table 4. Censorship by BMI Quartile



	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Trasnsitioned	0.2	24.1	26.8	38.66	30.4	66000.0
Censored	0.5	22.3	25.0	39.09	28.4	21920.0
All	12.1	23.9	26.7	27.49	30.2	74.5

Discrepancies

Table 4 shows some *very* concerning BMI values. The maximum ranges as high as 66000, and some as low as 10. A reasonable range for BMI values is between 18 and 60. We concluded that this was likely the result of an error during data entry or data transfer. Luckily, there were only 326 records that were outside of [12, 100], so we opted to remove those observations which were outside of that range. We decided on 100 as the upper limit because it *is* possible for a human to have a BMI close to 100. Furthurmore, our data is concerned with patients that already developed moderate hyptertension. So it is absolutely possible that an individual has a BMI between 60 and 100.

Next, we found that there *are* some redundant records for a single individual (multiple occurances of a single UNQID). This is because some individuals had their economic status re-evaluated during the study. Luckily, this accounted for only a small set of the data: 278 records. Our team decided to keep only the earliest records and drop any duplicate record with the re-evaluated socioeconomic status. Lastly, we found roughly 200 records with negative age values. Professor Ian Duncan, who had worked with the dataset in the past, advised us to remove all observations with either negative age values or survival times equal to zero. We followed his advisement.

Further Analysis

After removing observations with outlandish BMI values, duplicate observations, negative ages, and zero survival time, we continued our exploratory analysis. By this point in the project, our dataset contains 124684 rows.

Figure 1. Distribution of socioeconomic status by sex.

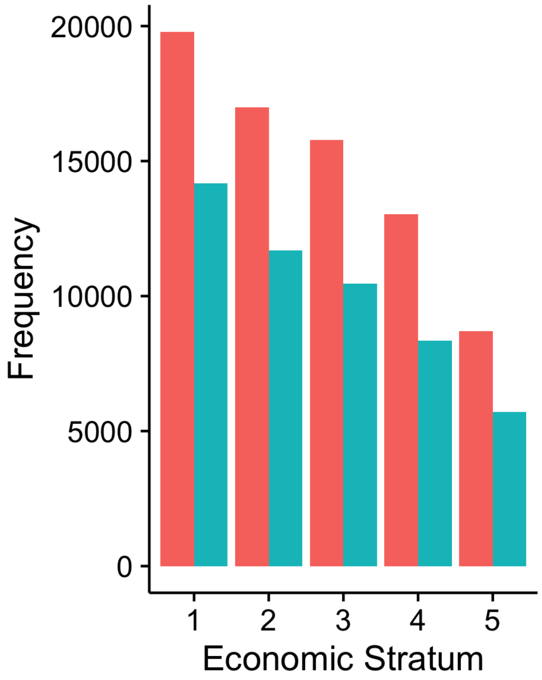
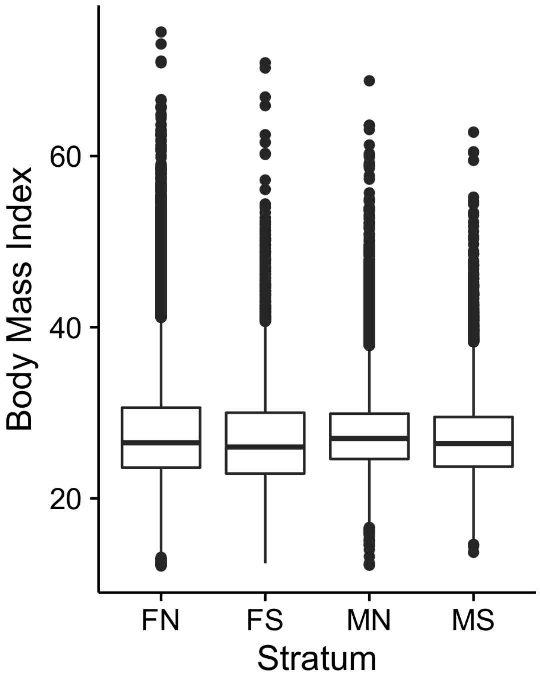


Figure 2. BMI distribution by sex and smoking status



In Figure 1, red corresponds to women and blue to men. We can see that there are more females in the dataset–roughly 3 women for every 2 men. The distributions of economic status are similar between the sexes. Figure 2 is a violin plot, so a wider horizontal portion corresponds to a more dense concentration of data at that value. It seems that women's BMI values are distributed more widely, while the men's BMI values are more densely distributed about their respective means.

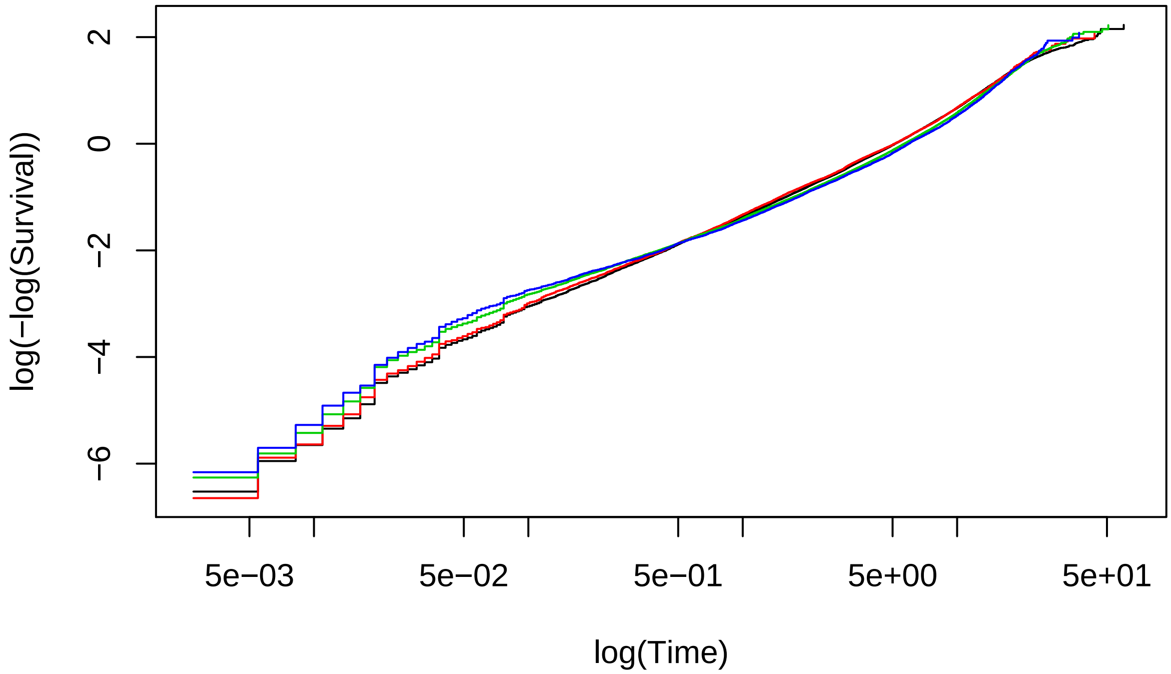
Modeling

Now that we have explored the dataset, we begin our modeling phase. First, our team constructed simple Kaplan-Meier estimate plots for each stratum of variables concerning alcohol consumption, Body Mass Index, sex, and cigarette consumption. We then moved on to construct two Cox Proportional-Hazards models: one with every covariate available, and a second with only those covariates which satisfied the proportional-hazards assumption. We constructed the reduced model to properly interpret the interactions. We constructed the full model in the hopes that it may provide some predictive power for future cases of sever hypertension. To test the predictive capabilities of the full model, we used cross validation on a single fold. All these steps are discussed in the following sub-sections.

Kaplan Meier Plots

Here we have Kaplan-Meier plots for each of the covariates listed in the section introduction.

Log−Log Plot of Sex and Smoking Status



Log−Log Plot of BMI Quartiles

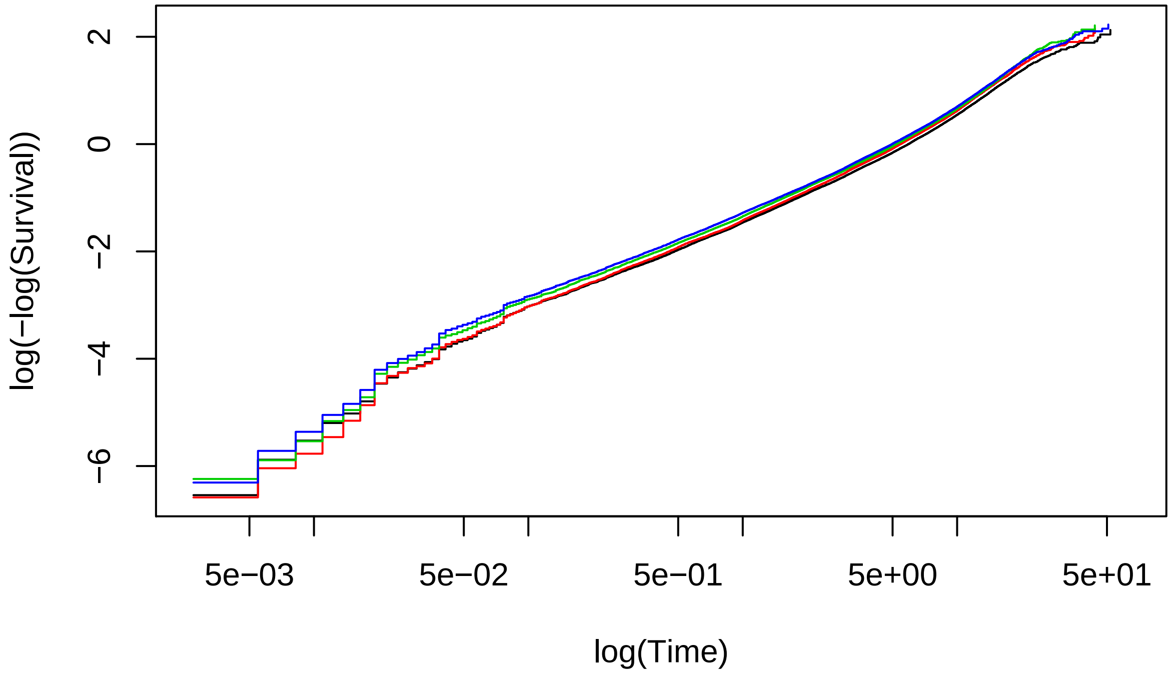


Figure 3. KM Estimates by Alcohol Consumption

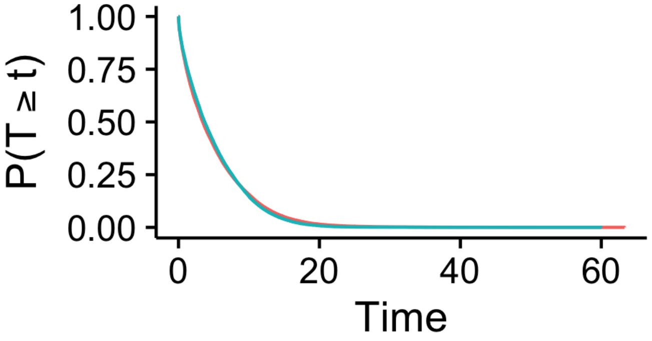


Figure 4. KM Estimates by BMI Level

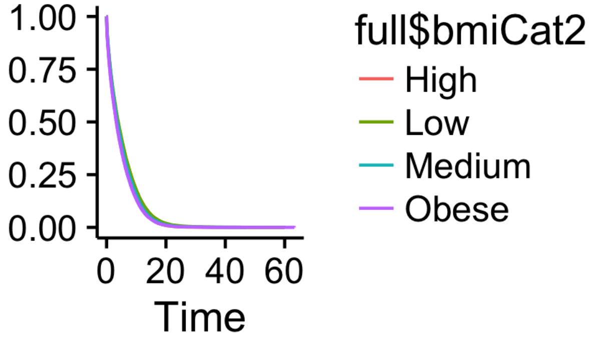
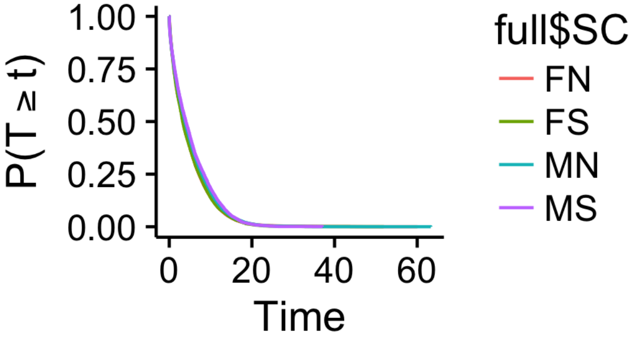


Figure 5. KM Estimates by Sex and Cigarette Usage



Fix these plots. Either one per row, or move legends into the open cell on bottom right (if that's possible).

The factor levels in Figure 4 are broken down by quartile; please see Table 5 shown below. In Figure 5, the first letter corresponds to the sex, and the second letter corresponds to their smoking status (**s**moker or **n**onsmoker).

Table 5. BMI Quartiles

Quartile Name	BMI Range
Low	[12.1, 23.9)
Medium	[23.9, 26.7)
High	[26.7, 30.2)
Obese	[30.2, 74.5]

Cox Modeling

We began our modeling with every available predictor. Then, after removing variables that did not satisfy the Proportional Hazards assumption, we arrived at a model using the patients' sex, age, cigarette usage, and BMI as the predictors. We tried to incorporate each patient's socio-economic status as an additional predictor, but we found that it failed the proportionality test.

Satisfying the Assumption

We began by finding a reduced model with significant predictors which entirely satisfied the proportionality assumption.

Table 6. Covatiates and Relevant Values

Covariate	e^{coef}	p-value for P.H. test
Age	1.0265404	0.1113
BMI	1.0224083	0.8120
Female Smoker	1.1066684	0.0808
Male Nonsmoker	0.9504585	0.4879
Male Smoker	1.0019740	0.2939

In Table 6, the covariates concerning sex and smoking status are all relative to female nonsmokers. For example, our model indicates that a male nonsmoker has an incident rate that is 95% of female nonsmokers (all else held equal). This combination of predictors is the only one that satisfied the proportional hazards assumption.

```
## Call:
## coxph(formula = obj ~ full$AGE_PRE + full$BMInew + full$SC)
## 
##   n= 124684, number of events= 111670 
## 
##                    coef  exp(coef)   se(coef)       z Pr(>|z|)    
## full$AGE_PRE  0.0262462  1.0265937  0.0002323 112.960  < 2e-16 ***
## full$BMInew   0.0225525  1.0228087  0.0005644  39.959  < 2e-16 ***
## full$SCFS     0.1114158  1.1178596  0.0109021  10.220  < 2e-16 ***
## full$SCMN    -0.0520974  0.9492364  0.0066694  -7.811 5.66e-15 ***
## full$SCMS     0.0098745  1.0099234  0.0121273   0.814    0.416    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
##              exp(coef) exp(-coef) lower .95 upper .95
## full$AGE_PRE    1.0266     0.9741    1.0261    1.0271
## full$BMInew     1.0228     0.9777    1.0217    1.0239
## full$SCFS       1.1179     0.8946    1.0942    1.1420
## full$SCMN       0.9492     1.0535    0.9369    0.9617
## full$SCMS       1.0099     0.9902    0.9862    1.0342
## 
## Concordance= 0.597  (se = 0.001 )
## Rsquare= 0.103   (max possible= 1 )
## Likelihood ratio test= 13533  on 5 df,   p=0
## Wald test            = 13264  on 5 df,   p=0
## Score (logrank) test = 13344  on 5 df,   p=0
```

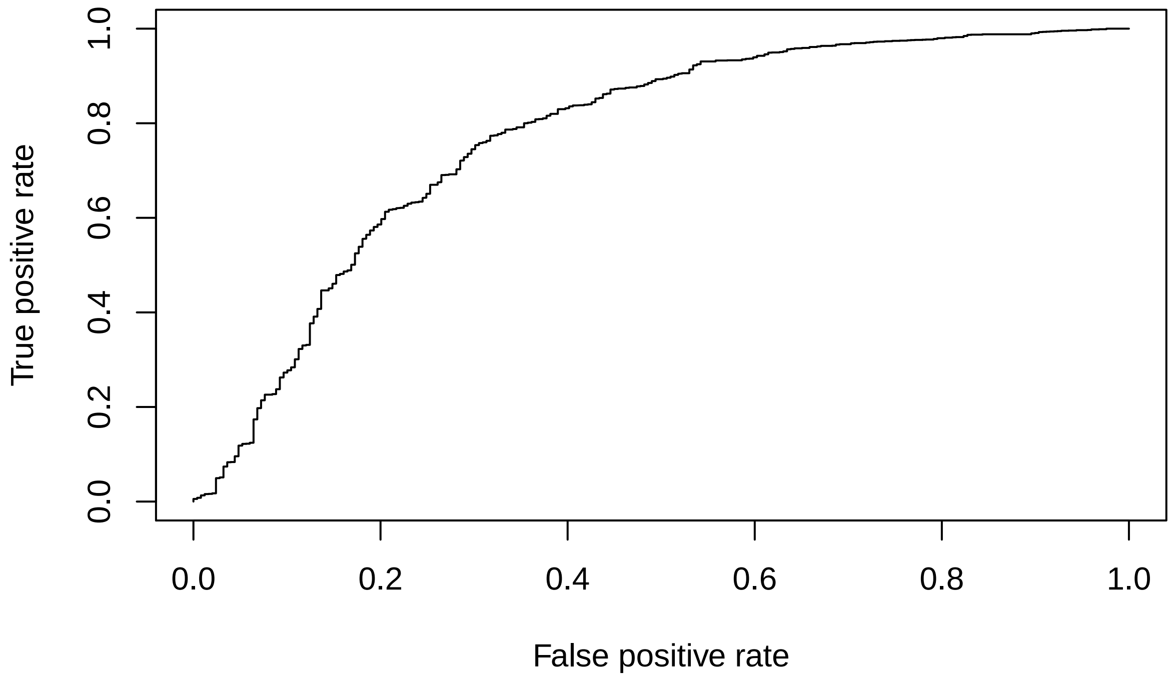
```
##                   rho  chisq      p
## full$AGE_PRE  0.00436  2.009 0.1563
## full$BMInew  -0.00157  0.264 0.6076
## full$SCFS    -0.00769  6.612 0.0101
## full$SCMN     0.00276  0.852 0.3559
## full$SCMS     0.00111  0.138 0.7105
## GLOBAL             NA 12.492 0.0286
```

Stratified Modeling

Now that we have built a model which satisfies the Cox Proportional Hazards Assumption, we will turn our attention to the socio-economic predictor. Our team was especially interested the effects of this predictor, so we decided to investigate it as a stratified covariate.

Cross Validation

```
## [1] TRUE
```



```
summary(logitModel)
```

```
##
## Call:
## glm(formula = event2 ~ AGE_PRE + BMInew + Socio.Economic + cigar +
##     sex, family = binomial(), data = train, control = control)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.5085   0.1937   0.2978   0.4415   1.7917
##
```

```
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -4.413423   0.339316 -13.007   <2e-16 ***
## AGE_PRE         0.082340   0.003440  23.933   <2e-16 ***
## BMInew          0.114944   0.008803  13.057   <2e-16 ***
## Socio.Economic -0.053096   0.030110  -1.763   0.0778 .
## cigar          -0.242542   0.098749  -2.456   0.0140 *
## sex            -0.056375   0.084667  -0.666   0.5055
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5168.5  on 8321  degrees of freedom
## Residual deviance: 4293.0  on 8316  degrees of freedom
## AIC: 4305
##
## Number of Fisher Scoring iterations: 6
```

Our model achieved an AUC score of 0.7719413 on the test set.

.

..

...

....

...

..

.

Acknowledgements

We would like to thank the following individuals for their contributions to the project. Some of our analysis would not have been possible without their assistance.

- Professor Ian Duncan and Shannon Nicponski
  - For advising our team throughout the project
- Terry M Therneau and Thomas Lumley
  - For authoring and maintaining the Survival package
- Tal Galili
  - For publicizing his ggsurv function on R-statistics.com
- Hadley Wickham
  - For authoring the plyr, dplyr, and ggplot2 packages