Predictive Modeling Workshop

Pre-Work: Data Cleaning / Exploration and Feature Engineering

March 21, 2018

ILTCI

Important Info

- Claim termination rate (CTR) data has been extracted from the SOA experience database[1] and modified for workshop purposes. Don’t use the modified data in a real study.
- The modified data can be found here. The R programs referenced within this presentation can be found here.
- This data will be used with the pre-work materials and in the workshop as well.

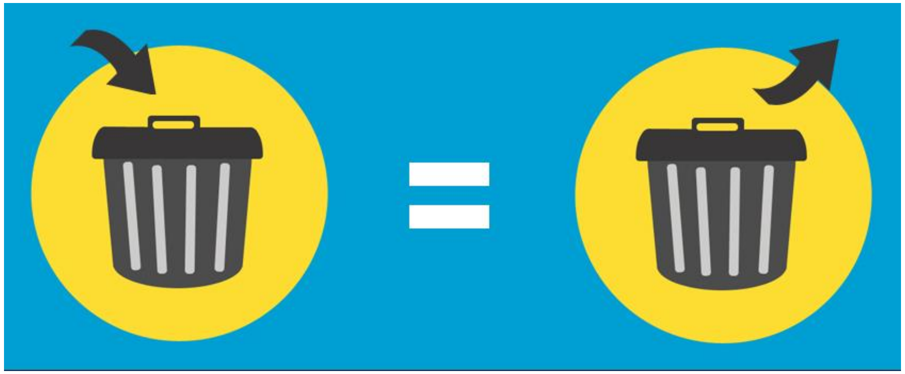
[1] https://www.soa.org/experience-studies/2015/research-ltc-study-2000-11-aggregrated/

Directories and Software

- On your computer make sure the following directories exist, which are needed to run the programs referenced in this presentation:
  - Create C:\ILTCI_Workshop\Data\
  - Make sure you put the CTR data in the above directory
  - Also create C:\ILTCI_Workshop\Output\
- On your computer you will want to be sure that you have installed the core R software and R Studio. Please go to these links if not:
  - Download the core R software here.
  - Download RStudio here.

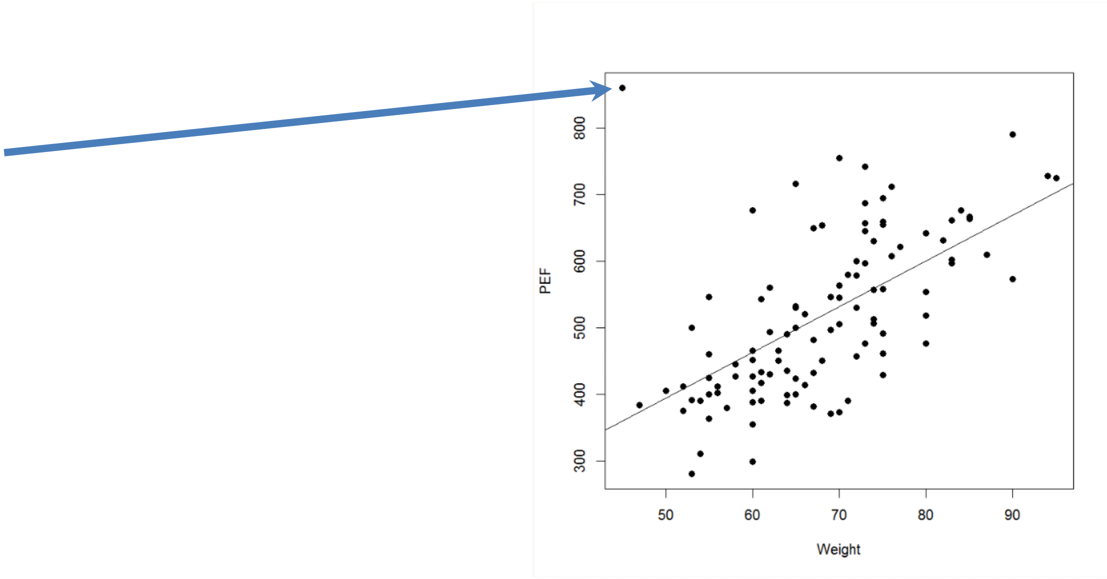
Overview

- Now we have data let’s start modeling right? Nope….
- Garbage in is garbage out, make sure your data is in good shape
- ~80% of your time will be spent here



Getting started

- It’s important to make sure your data are clean and ready to go for your modeling project.
- Have you viewed your driver and response variables?
- Does your data have any outliers, or blanks?



Source: http://stats.stackexchange.com/questions/194783/extreme-values-in-the-data

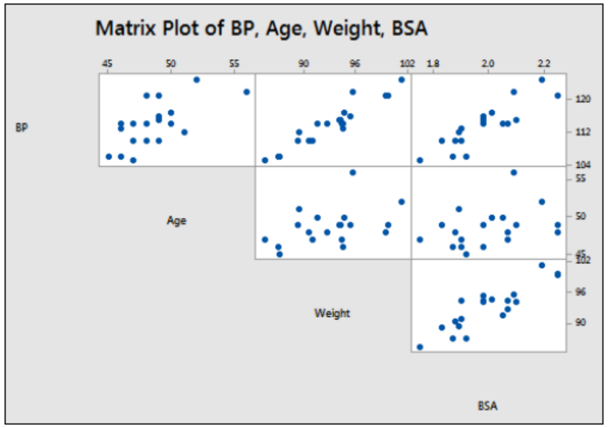
Getting started

- Now let’s jump into some R code. Open up the “0100 - Data Explore.R” program and go through steps 1.0.1 to 1.0.7

- Make sure that you have already run R program 0001 – Install Packages, which was attached to the homework reminder e-mail.

Data relationships

- If you notice patterns in your data that are strange find out why and try and fix them.
- Have you viewed relationships amongst your driver and response variables?
- Are some driver variables so related that they do not appear to add any value?



Source: https://onlinecourses.science.psu.edu/stat501/node/347

Moment of truth

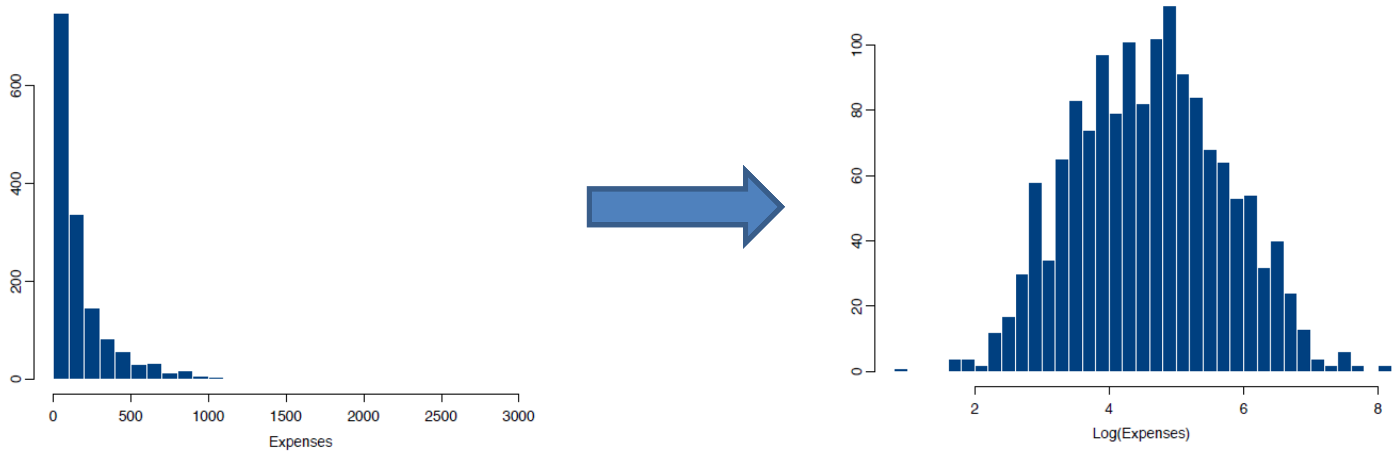
- Sometimes your data may not be good enough quality to use in a modeling project.
- Are the data granular enough for your modeling purpose?
- Are the data credible enough? Can industry data help out?
- Do you find too many outliers, blanks, and / or other suspicious patterns?
- Do you find that certain relationships you expect to be true, are violated by simple views of your data?

Moment of truth

- Jump back into the “0100 - Data Explore.R” program and go through steps 1.0.8 to 1.0.13

Variable transformations

- Data is clean – Can we make any additional variables?
- Start simple – We can always circle back.
- Would any transformations of your variables better serve you?



Source: http://www.kenbenoit.net/courses/ME104/logmodels2.pdf

Base level characteristics

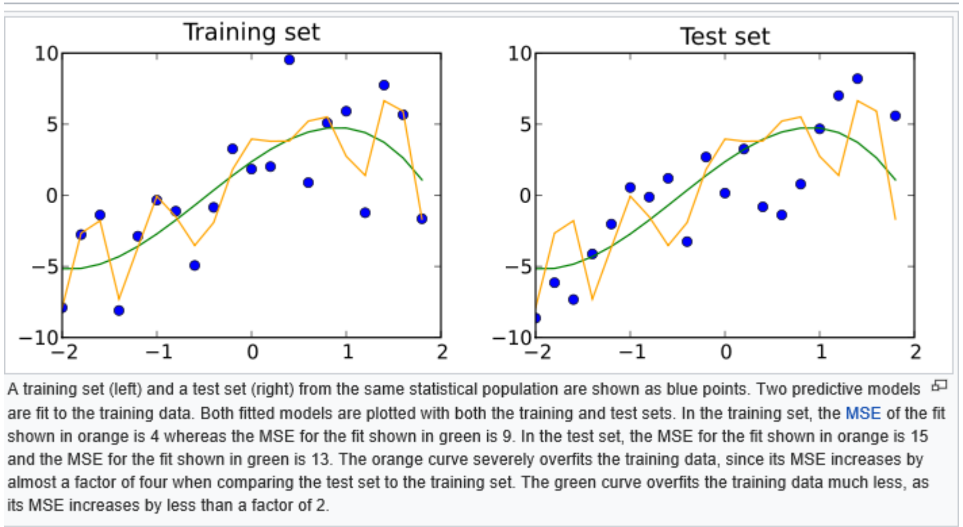
- The most populous part of your data is what can be considered the “base level” characteristic.
- For example, if there are more claim terminations for females than males, then females would be the desirable base level characteristic of the gender driver variable.
- This concept will become more important in the next section.

“The base level should not be sparse. … Any level which is not sparse is an appropriate base level.”

– Piet de Jong and Gillian Z. Heller, “Generalized Linear Models for Insurance Data”

Splitting the datasets

- Split data into training and validation data.
- What proportion of your data can you afford to segment out of your training data analysis?
- Can you afford to also make a test dataset?



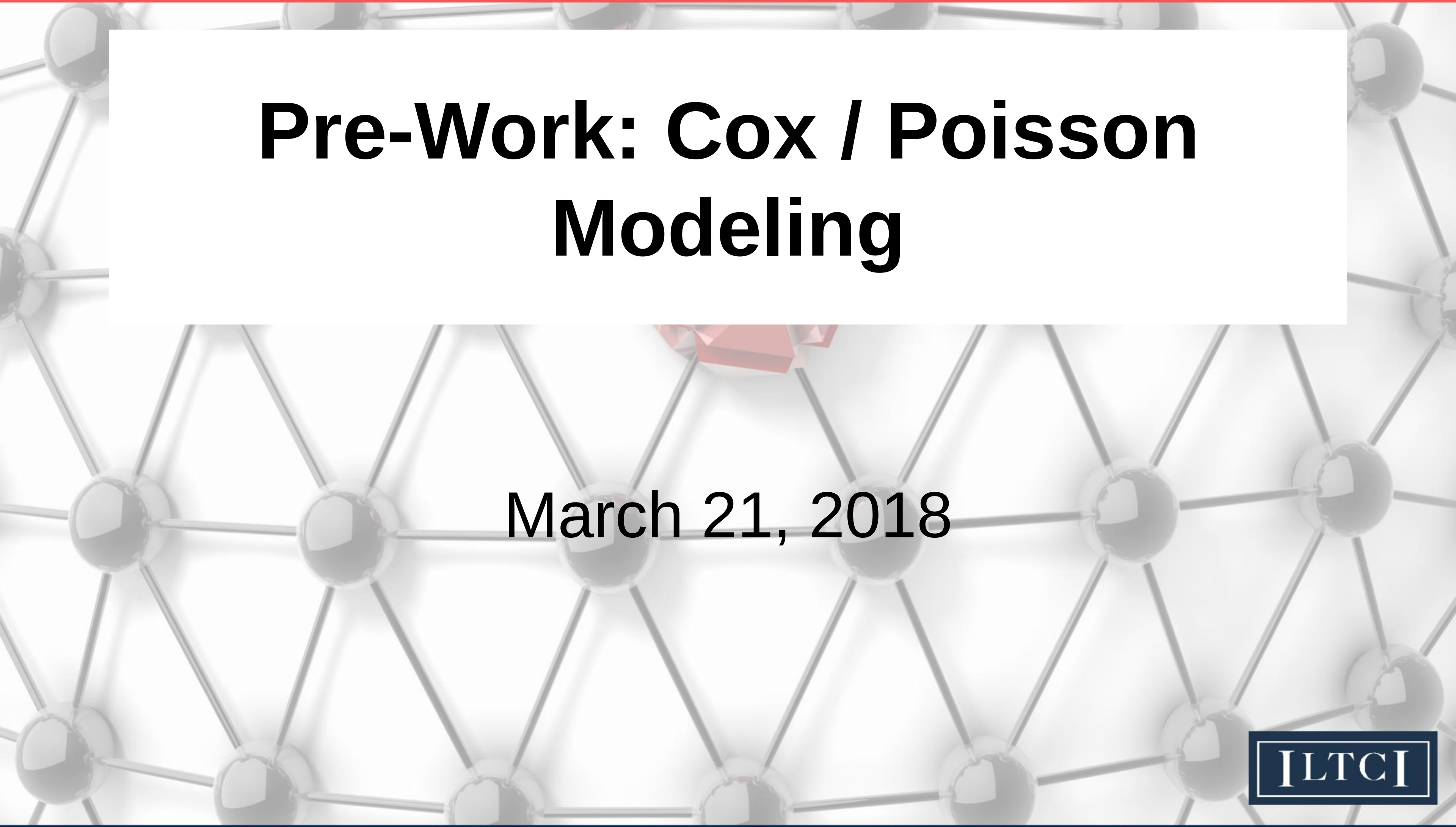
A training set (left) and a test set (right) from the same statistical population are shown as blue points. Two predictive models are fit to the training data. Both fitted models are plotted with both the training and test sets. In the training set, the MSE of the fit shown in orange is 4 whereas the MSE for the fit shown in green is 9. In the test set, the MSE for the fit shown in orange is 15 and the MSE for the fit shown in green is 13. The orange curve severely overfits the training data, since its MSE increases by almost a factor of four when comparing the test set to the training set. The green curve overfits the training data much less, as its MSE increases by less than a factor of 2.

Source: https://en.wikipedia.org/wiki/Test_set

Splitting the datasets

- Let’s now split the data in R using the following program “0150 - Separate data into train val test.R”

Predictive Modeling Workshop



Pre-Work: Cox / Poisson Modeling

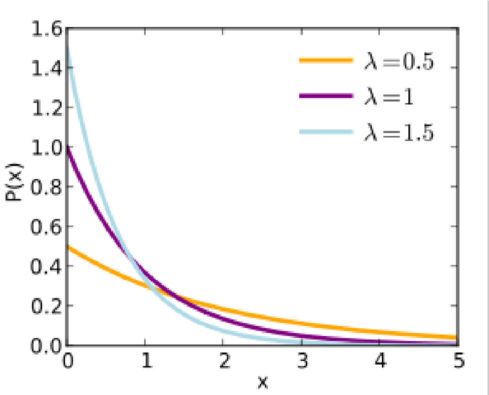
March 21, 2018

Overview

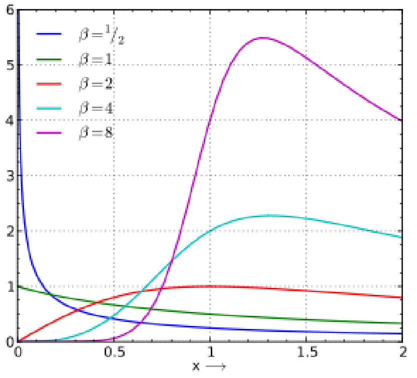
- We explored and staged the data for modeling, now let’s review the basics of survival modeling and how they can be applied with predictive modeling.
- For more background on survival modeling review items #4 and #5 of the pre-workshop materials from the 2017 ILTCI workshop located here: http://iltciconf.org/2017/predictivemodelingmaterials.htm

Overview

- Start simple to get a feel for your data.
- Do you have some *a priori* notions as to what relationships should exist in your dataset?
- Do you have a sense as to how the hazard rate function might look?
- R can help you fit models based on your answers to the above.



Exponential failure density functions. Each of these has a (different) constant hazard function (see text).



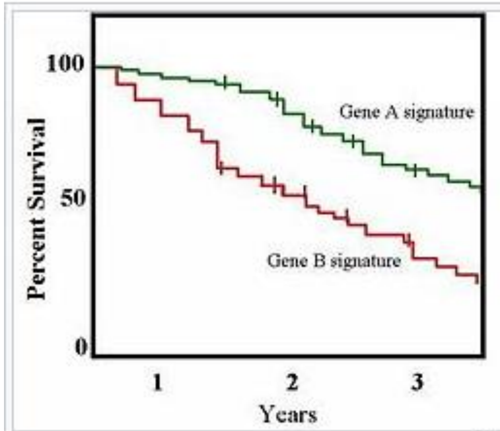
Hazard function $h(t)$ plotted for a selection of log-logistic distributions.

Source: https://en.wikipedia.org/wiki/Failure_rate

Survival functions

- Kaplan-Meier produces a survival function based on your raw input data.
- What will you use as your function of time?
- What cells will you focus on?

$$S(t) = \prod_{t=1}^{n} \left(1 - \frac{d_t}{n_t}\right)$$

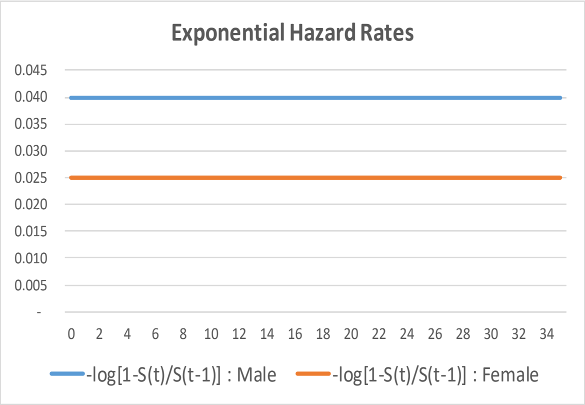


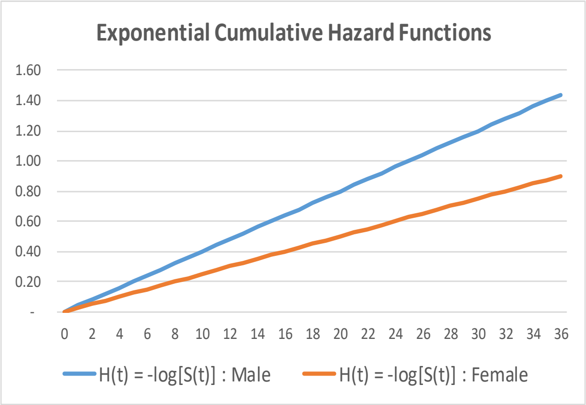
An example of a Kaplan–Meier plot for two conditions associated with patient survival.

Source: https://en.wikipedia.org/wiki/Kaplan%E2%80%93Meier_estimator

Hazard functions

- A Nelson-Aalen estimator can give you an idea about the shape of the cumulative hazard function, and as a result which model of hazard rates might best fit your data.
- Which probability distribution might best describe the CTR data?



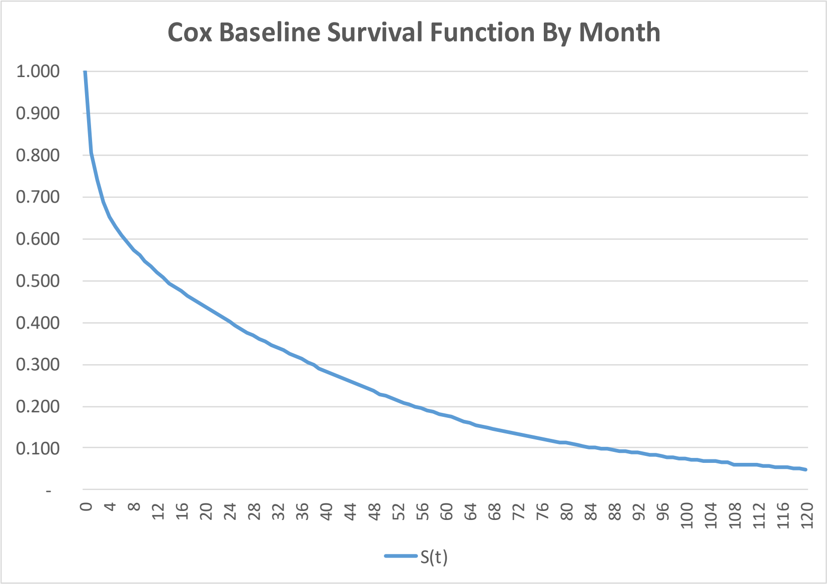


Hazard functions

- Now let’s try out code that explores some basic tabular survival curves in the “0200 - KM-NA.R” program.

The Cox proportional hazard model

- The Cox model produces a non-parametric baseline hazard rate function $\lambda(t)$ or $h(t)$.
- It also produces an estimate of the regression coefficients for covariates like gender and their statistical significance.



The Cox proportional hazard model

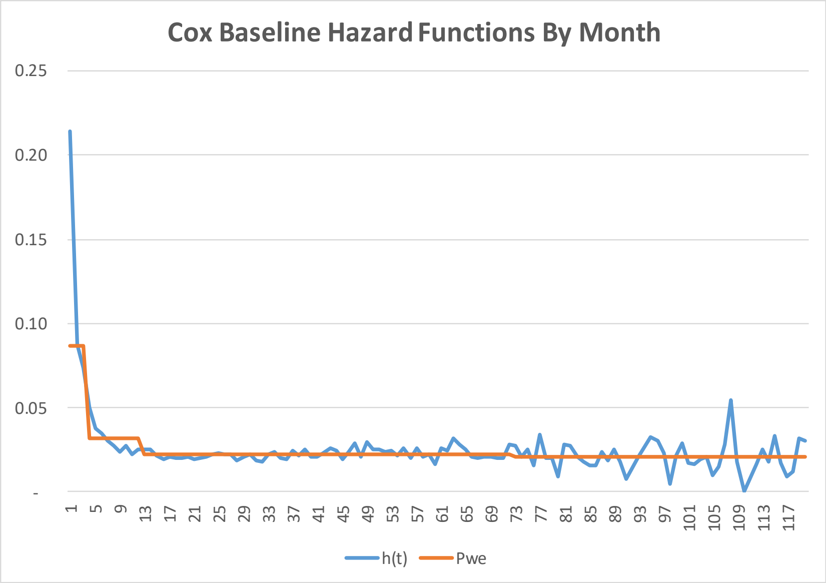
- In the “0220 - CoxPH.R” program we will now show how we can use a Cox model to gain more insight from our data by analyzing it all at once to find relationships across cohorts.

Interpreting the Cox results

- The plot shows the base level survival function.
- Covariate values show how the CTRs vary for the non-baseline effects.
- Tests of statistical significance usually follow a Chi-Square test.
- Effects are relative to the base level – say for gender, female might be base level (with subscript “0”), so male would be an effect that is:
  - Exponential in $S(t)$: $S(t) = S_0(t)^{\exp(\mathit{Effect})}$
  - Proportional in $h(t)$: $\exp(\mathit{Effect}) \times h_0(t)$
  - Linear in $\log[h(t)]$: $\log[h(t)] = \mathit{Effect} + \log[h_0(t)]$

Comparable models

- A piecewise Exponential model can be fit to the data and provide very similar results to a Cox model with more time granularity.
- The piecewise Exponential can be fit to about four segments of the claim duration curve and arrive at a similar result.

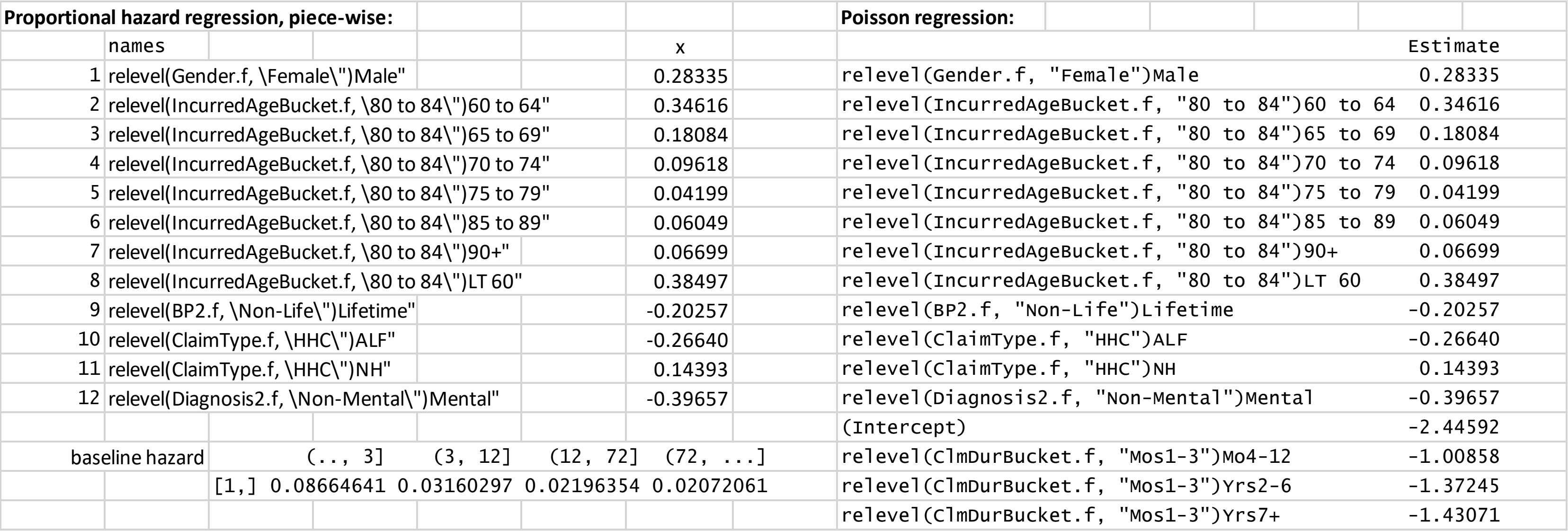


Comparable models

- In steps 2.3.1 to 2.3.9 of the “0230 - PweCoxPH.R” program we will explore running a piece-wise exponential model, which can allow us to introduce time-varying covariates.

Poisson-Exponential model equivalence

- A process that has Poisson counts is equivalent to a process that has Exponential waiting times.
- The Poisson GLM can be used to arrive at the exact same results as with the piecewise Exponential model.



Proportional hazard regression, piece-wise:			Poisson regression:	
	names	x		Estimate
1	relevel(Gender.f, \Female\")Male"	0.28335	relevel(Gender.f, "Female")Male	0.28335
2	relevel(IncurredAgeBucket.f, \80 to 84\")60 to 64"	0.34616	relevel(IncurredAgeBucket.f, "80 to 84")60 to 64	0.34616
3	relevel(IncurredAgeBucket.f, \80 to 84\")65 to 69"	0.18084	relevel(IncurredAgeBucket.f, "80 to 84")65 to 69	0.18084
4	relevel(IncurredAgeBucket.f, \80 to 84\")70 to 74"	0.09618	relevel(IncurredAgeBucket.f, "80 to 84")70 to 74	0.09618
5	relevel(IncurredAgeBucket.f, \80 to 84\")75 to 79"	0.04199	relevel(IncurredAgeBucket.f, "80 to 84")75 to 79	0.04199
6	relevel(IncurredAgeBucket.f, \80 to 84\")85 to 89"	0.06049	relevel(IncurredAgeBucket.f, "80 to 84")85 to 89	0.06049
7	relevel(IncurredAgeBucket.f, \80 to 84\")90+"	0.06699	relevel(IncurredAgeBucket.f, "80 to 84")90+	0.06699
8	relevel(IncurredAgeBucket.f, \80 to 84\")LT 60"	0.38497	relevel(IncurredAgeBucket.f, "80 to 84")LT 60	0.38497
9	relevel(BP2.f, \Non-Life\")Lifetime"	-0.20257	relevel(BP2.f, "Non-Life")Lifetime	-0.20257
10	relevel(ClaimType.f, \HHC\")ALF"	-0.26640	relevel(ClaimType.f, "HHC")ALF	-0.26640
11	relevel(ClaimType.f, \HHC\")NH"	0.14393	relevel(ClaimType.f, "HHC")NH	0.14393
12	relevel(Diagnosis2.f, \Non-Mental\")Mental"	-0.39657	relevel(Diagnosis2.f, "Non-Mental")Mental	-0.39657
			(Intercept)	-2.44592
baseline hazard	(.., 3] (3, 12] (12, 72] (72, ...]		relevel(ClmDurBucket.f, "Mos1-3")Mo4-12	-1.00858
	[1,] 0.08664641 0.03160297 0.02196354 0.02072061		relevel(ClmDurBucket.f, "Mos1-3")Yrs2-6	-1.37245
			relevel(ClmDurBucket.f, "Mos1-3")Yrs7+	-1.43071

Poisson-Exponential model equivalence

- Finally in steps 2.3.10 to 2.3.14 of the “0230 - PweCoxPH.R” program we will run through code that shows the equivalence of a piece-wise exponential model and a Poisson GLM.

Summary

- We have introduced the data and provided a link between survival models and their application through GLMs.
- Important notes:
  - When using a Poisson GLM for modeling hazard rates the exact exposure method should be used to calculate exposures versus using the actuarial exposure method. The exposures should also be used as an offset in the model.
  - The resulting predictions of the model are then hazard rates. To convert them into probabilities take:
    1- exp(-hazard rate)

Items of note

- Not all of the files you downloaded were used in this pre-work presentation.
- We have not used all of the R programs in the downloaded zip file. The others will be used in the workshop.
- There is an additional data file called “smoothed_assumptions.Rdata”. That file will not be used until the workshop as well.