Survival Analysis of Worsening Hypertension

PSTAT 196: Research in Actuarial Science

Spring Quarter, 2017

Outline

1. About the Dataset
2. Background
3. Objective and methods
4. Data Cleaning
5. Exploratory Analysis
6. Modeling
    a. Cox P.H.
    b. Cross Validation
    c. Interpretation
    d. Parametric Modeling
7. Second Dataset
    a. New Variables
    b. Difficulties
    c. Interpretation
8. Future Directions

Undergraduate Researchers
Jason Freeberg
Ziyi (Terry) Jiang
John Randazzo

Advisors
Ian Duncan
Shannon Nicponski

About the Dataset

- An extract from database THIN (The Health Improvement Network) based in England and Wales.
- Collected from mid-1990's through 2012
- Limitations:
  - 1 observation per person
  - Erroneous data entry - have to assume correctness aside from impossible values
- Everyone already has moderate hypertension (stage 2 of 3)
  - Transition from stage 2 to 3 is the event of interest for our study

Background

Hypertension (a.k.a. high blood pressure)

- About 75 million American adults (29%) have high blood pressure.[1]
- Total costs associated with high blood pressure in 2011 in the US were $46 billion.[1]
- Two stages of Hypertension
  - Primary → 90% of all cases, nonspecific lifestyle or genetic causes
  - Secondary → Attributed to specific diseases or disorders

[1] "High Blood Pressure Frequently Asked Questions (FAQs)," *Centers for Disease Control and Prevention*, accessed May 30, 2017.

Objectives and Methods

- Identify factors leading to severe hypertension
    - Using a Cox Proportional-Hazards Model
    - In the original dataset
- Further test our Cox P.H. model
    - Using Cross-Validation
- Investigate the medical factors leading to severe hypertension
    - In the 2nd dataset
- Extend the Cox model to time-dependent covariates
    - Through stratification and parametric models

Variables of Interest

Name	Description	Support
AGE_PRE	Age in years when subject entered the study	[-55, 101]
alcohol	Binary indicator for the use of alcohol (0: no, 1: yes)	{0,1}
BMInew	BMI at start of study in kg/m^2	[0.2, 66000]
Cigar	Binary indicator for the use of cigarettes (0: no, 1: yes)	{0, 1}
DURATION	Number of **days** between entering study and transitioning	[0, 41180]
Partial_code_ff	Transition code (2000: moderate hypertension, 3000: severe hypertension)	{2000, 3000} *{0,1}
Sex	Binary indicator for sex (1: male, 2: female)	{1, 2}
Socio-Economic	Categorical variable ranging from affluent (1) to deprived (5)	{1, 2, 3, 4, 5}

Data Cleaning and Vetting

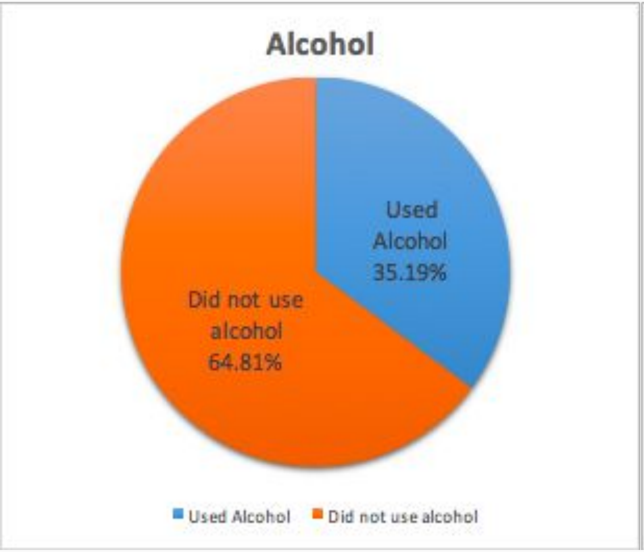
- Removed all subjects with:
  - BMI < 12 or BMI > 100
  - Starting age <= 0
  - Duration time equal to 0
  - Duplicate observations (no differences between the two)
- Define new categorical predictor: SC = sex * cigar
  - 4 factor variables: Male smoker, Male non-smoker, Female smoker, Female non-smoker
  - MS, MN, FS, FN

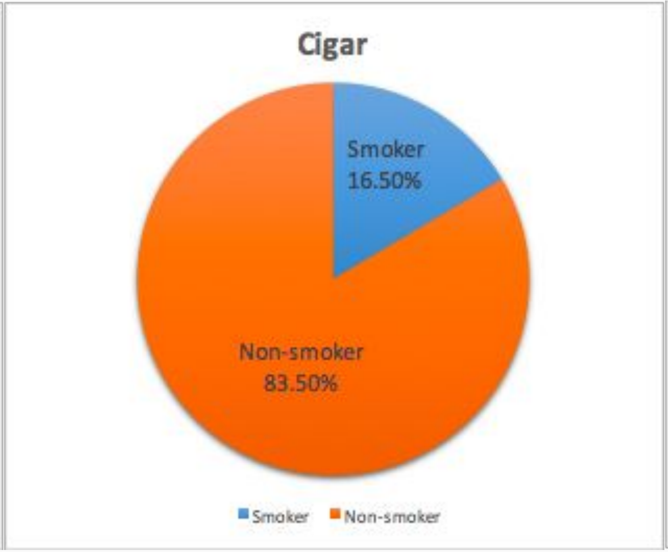
Dataset post-cleaning

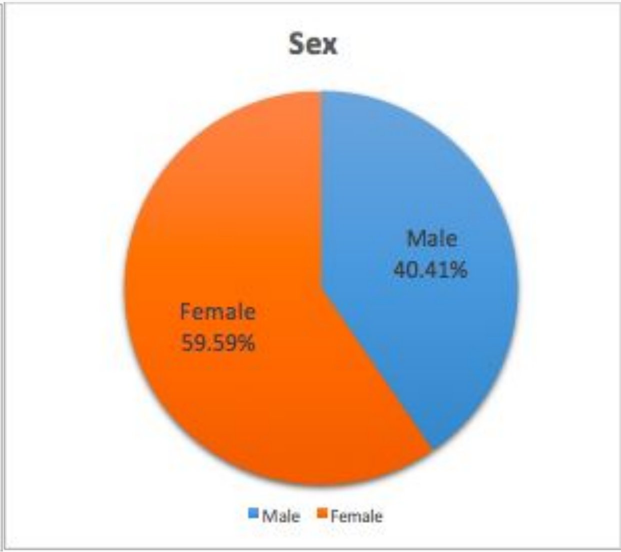
Variable	Before Cleaning	After Cleaning
AGE_PRE	[-55, 101]	[8, 101]
BMInew	[0.2, 66000]	[12.1, 98.5]
DURATION (days)	[0, 41180]	(0, 41180]

- The number of observations shrinks from 126,665 to 124,393
    - 1.79% removed

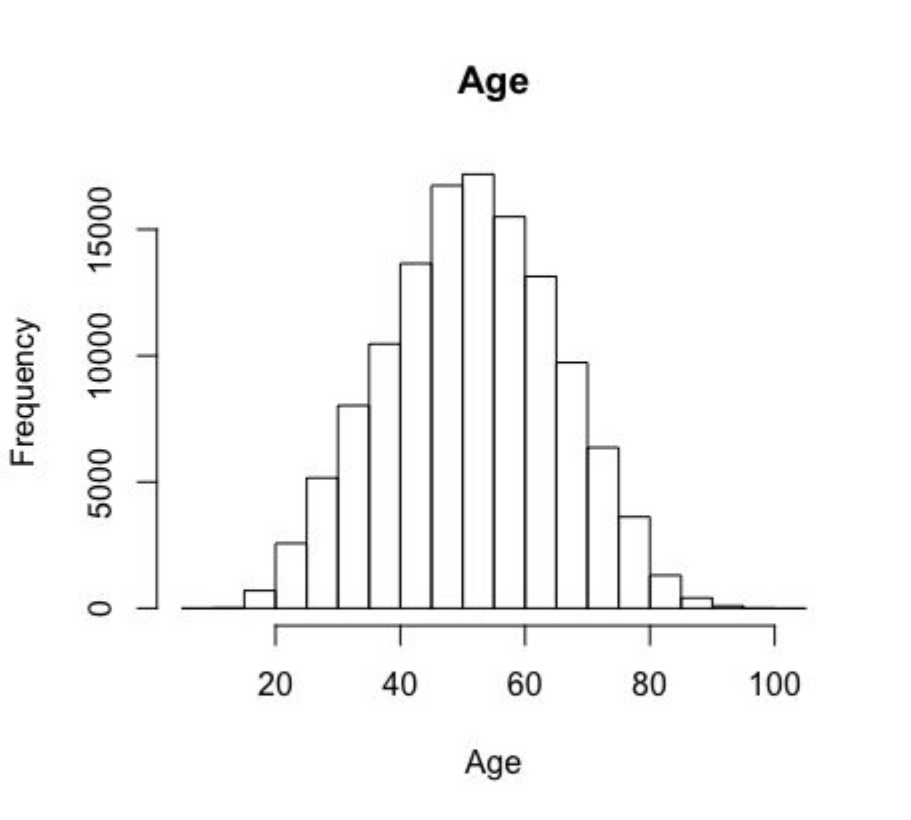
Exploratory Analysis



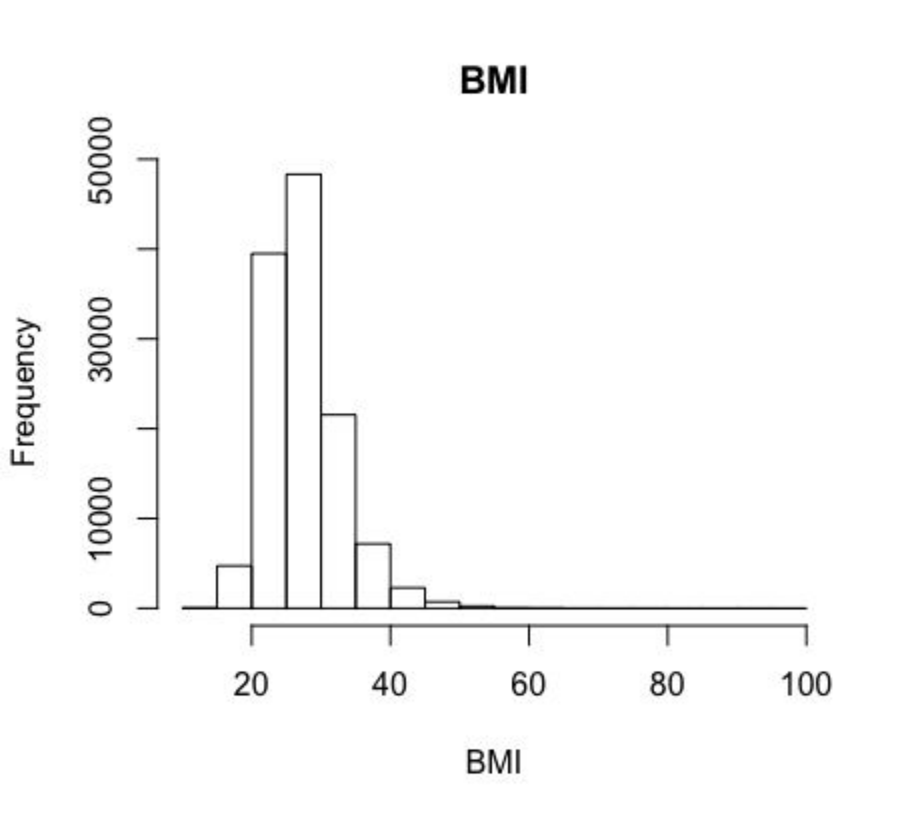




Exploratory Analysis

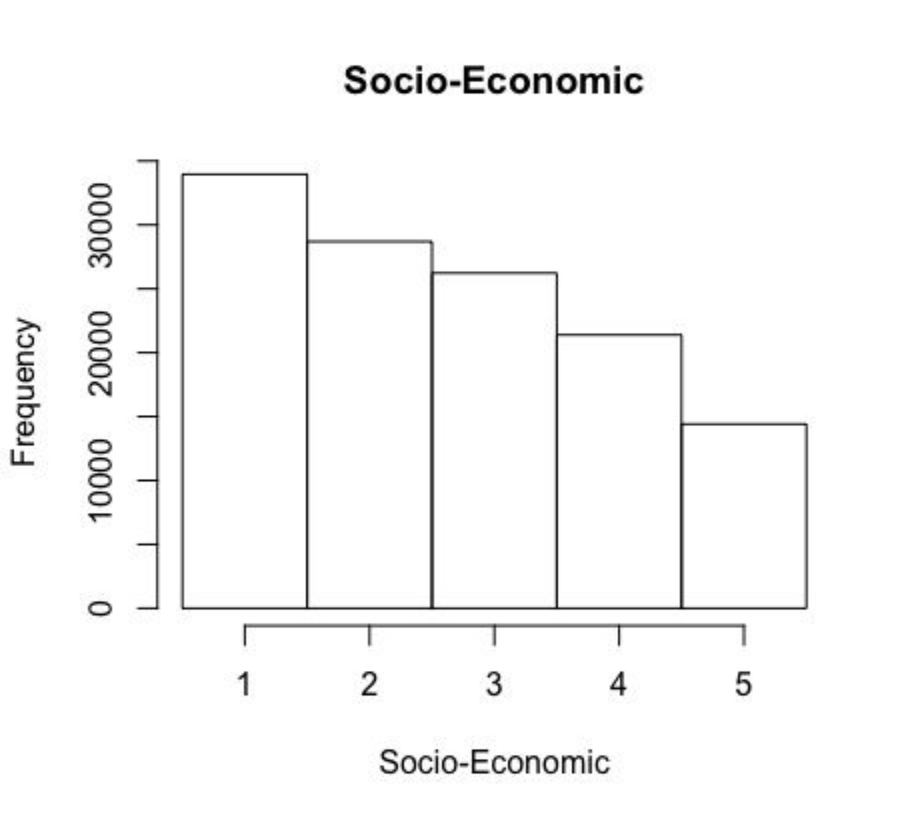


Mean = 51.90



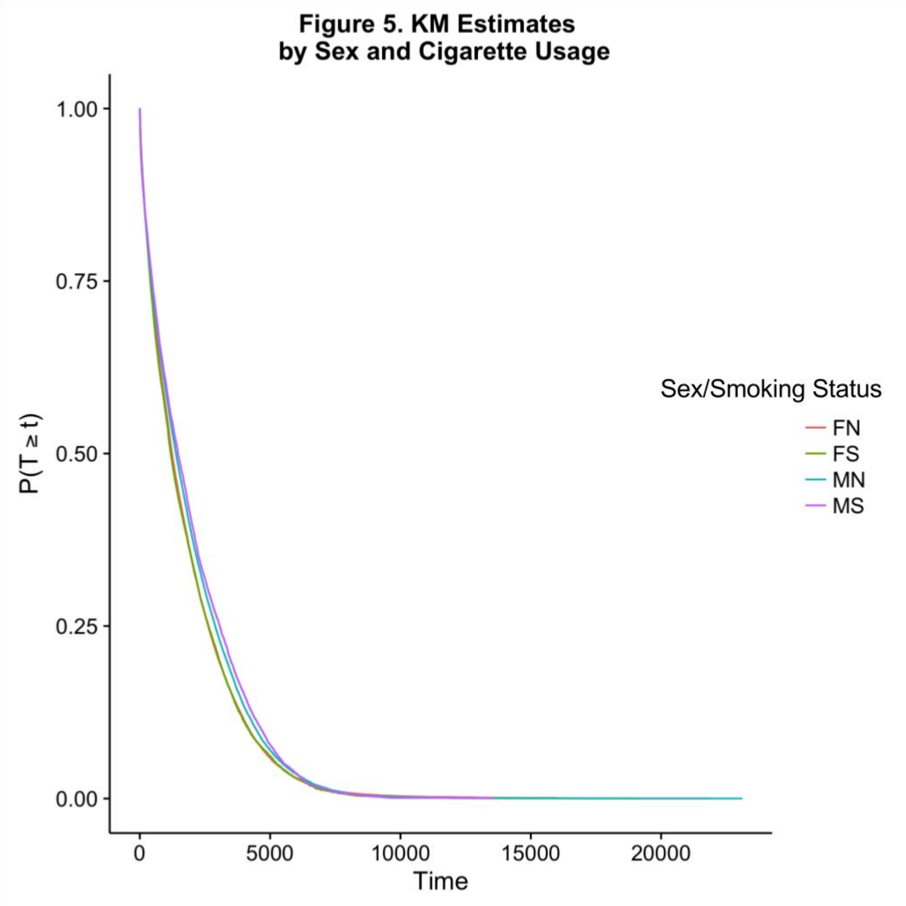
Mean = 27.50

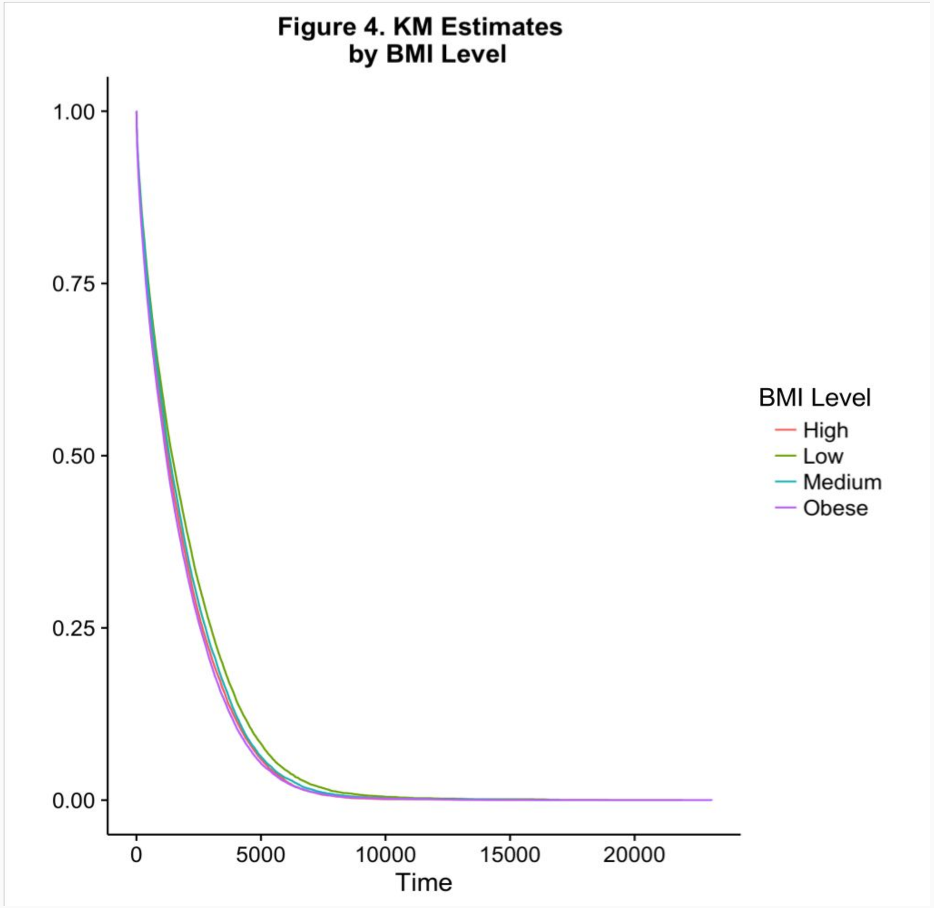
Exploratory Analysis



Socio-Economic	Counts	% of population
1 (affluent)	33959	27.23%
2	28697	23.01%
3	26231	21.04%
4	21391	17.16%
5 (deprived)	14414	11.56%

Exploratory Modeling





Cox Proportional Hazards Model Review

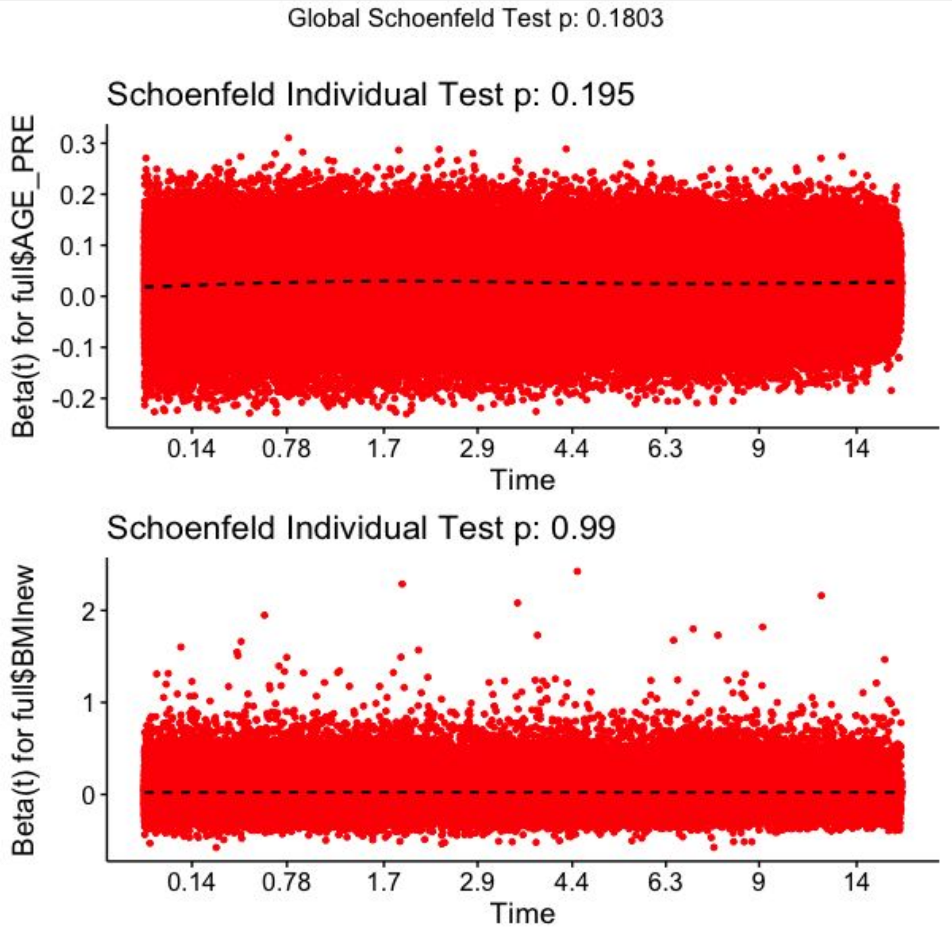
$$h(t) = h_0(t)\exp(\beta_1 X_1 + \cdots + \beta_n X_n)$$

- The Cox approach models the hazard rate, $h(t)$, over time
- Increase in a covariate → multiplicative effect on $h(t)$
- The baseline hazard, $h_0(t)$, is **not** estimated
- So the Cox model assumes that the covariates' effects are multiplicatively (or proportionally) related to $h(t)$

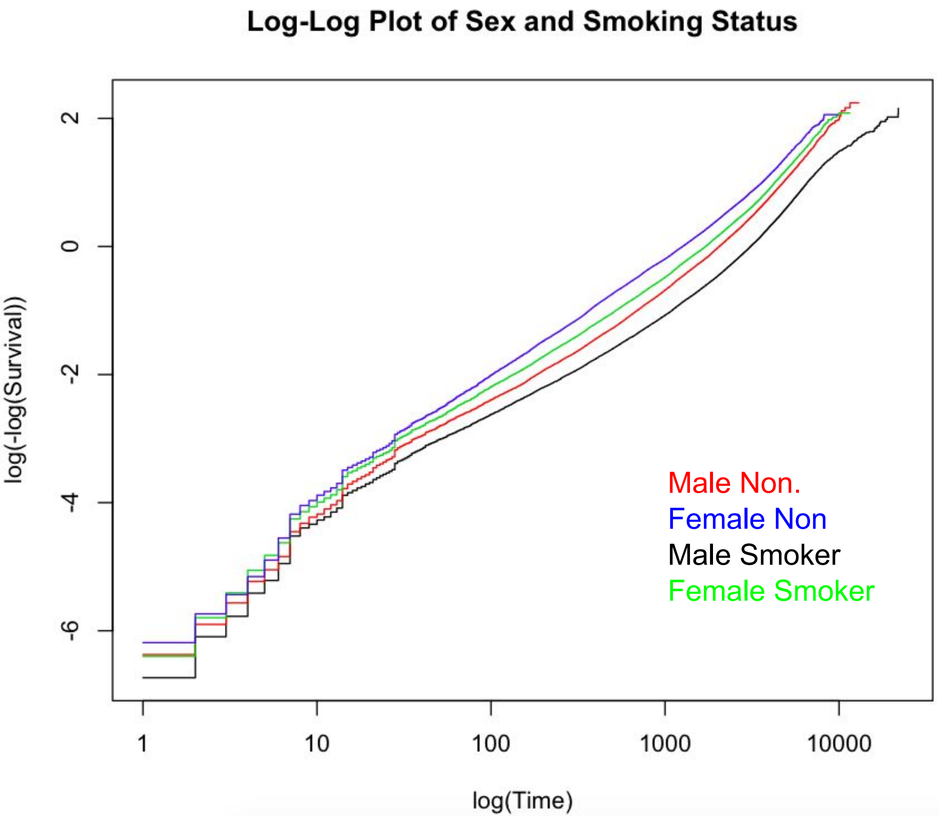
Checking the Model Assumptions

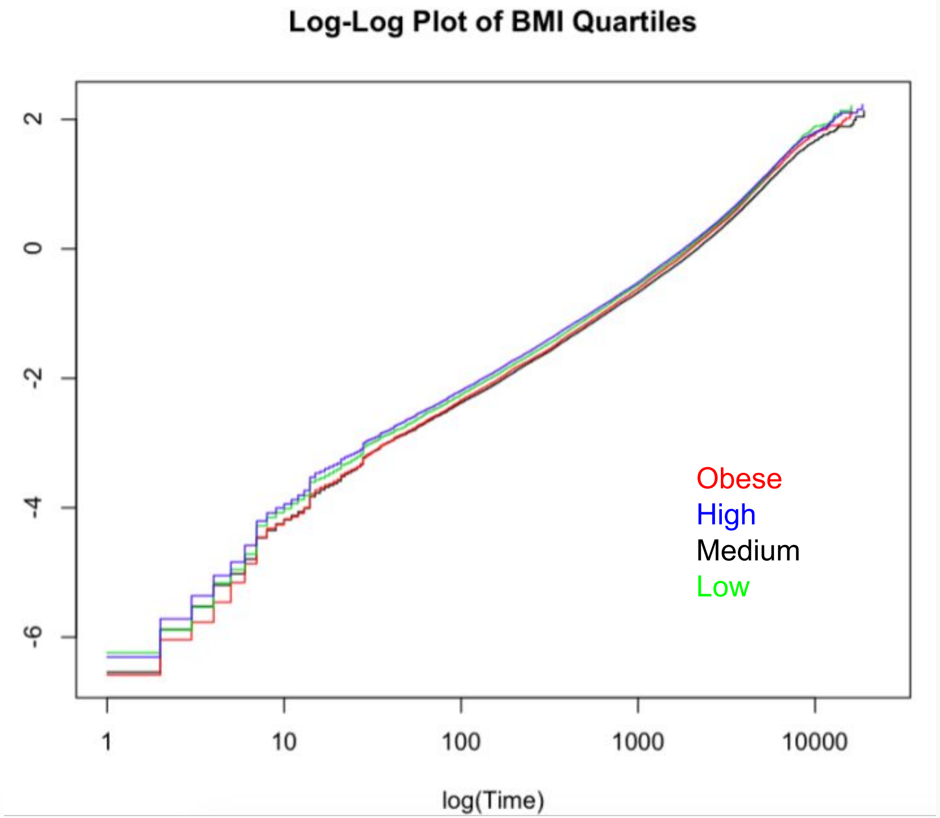
- Schoenfeld Residual Plots for continuous covariates
  - H_0 : $\mathit{PH\ Assumption\ is\ met}$
  - H_α : $\mathit{PH\ Assumption\ is\ not\ met}$
  - High p-values imply horizontal line across entire time span → agrees with assumption
- Log-log plots for categorical/discrete covariates
  - Schoenfeld Residuals yield no insight about alignment with PH assumption with regards to categorical predictors
  - Transforming Kaplan-Meier Estimator (x = Time, y = Survival Prob.) into Log-log plot (x = log(Time), y = log(-log(Survival Prob.))) can help us assess assumption
  - Parallel lines for each risk group implies assumption is met

Schoenfeld Residuals



Log(-Log(Survival)) Plots





Our Stratified Cox Model

- Specifically, socio-economic class did not meet PH assumption when tested numerically
  - Anticipated this to be an important factor → stratified the covariate
- There are 5 different levels of socio-economic class, a Cox model is fit for each level
- The covariates/regression coefficients are **the same** for all 5 strata
  - Only the baseline hazard rate differs across the strata

Our Stratified Cox PH Model

$$h(t) = h_i(t)\exp(0.0261 \cdot age + 0.0216 \cdot bmi + 0.1008 \cdot FS + 0.0015 \cdot MS + 0.0505 \cdot MN) \quad i = 1,2,3,4,5$$

Used the proportionality tests and BIC as our guiding criteria ($BIC = ln(n)k - 2ln(L)$)

Covariate	Estimated Hazard Rate	95% Confidence Interval
Age	1.0265	[1.0260, 1.0269]
BMI	1.022	[1.0206, 1.0228]
Female Smoker	1.106	[1.0826, 1.1301]
Male Non-Smoker	0.950	[.9384, .9632]
Male Smoker	1.0019	[.9779, 1.0256]

Baseline Hazard Rates for Different Socio-Economic Strata

Time = 1 year

Socio Economic	Hazard Rate
1 (affluent)	0.2281420
2	0.2263601
3	0.2409130
4	0.2549135
5 (deprived)	0.2707321

Time = 3 years

Socio Economic	Hazard Rate
1	0.5564318
2	0.5557649
3	0.5792660
4	0.6067651
5	0.6391335

Time = 5 years

Socio Economic	Hazard Rate
1	0.8963512
2	0.8929580
3	0.9307331
4	0.9454276
5	0.9947118

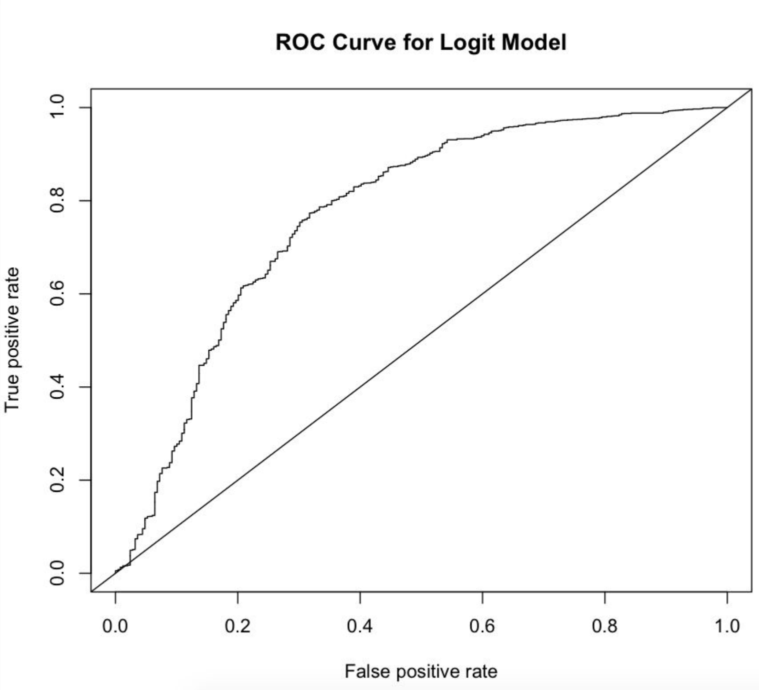
Cross-Validation

Like regression and classification, we wanted to see how robust our Cox model is

- Difficult because we are handling **duration** and **state**
  - First tried to predict duration
- Moved to predicting **state** at 3.5 years
- But the Cox model does not estimate the baseline rate!
  - Tried using the following approximation: $\hat{h}_0(t) = \sum_{t<t^*} h(t)$
  - … With poor results

Cross-Validation

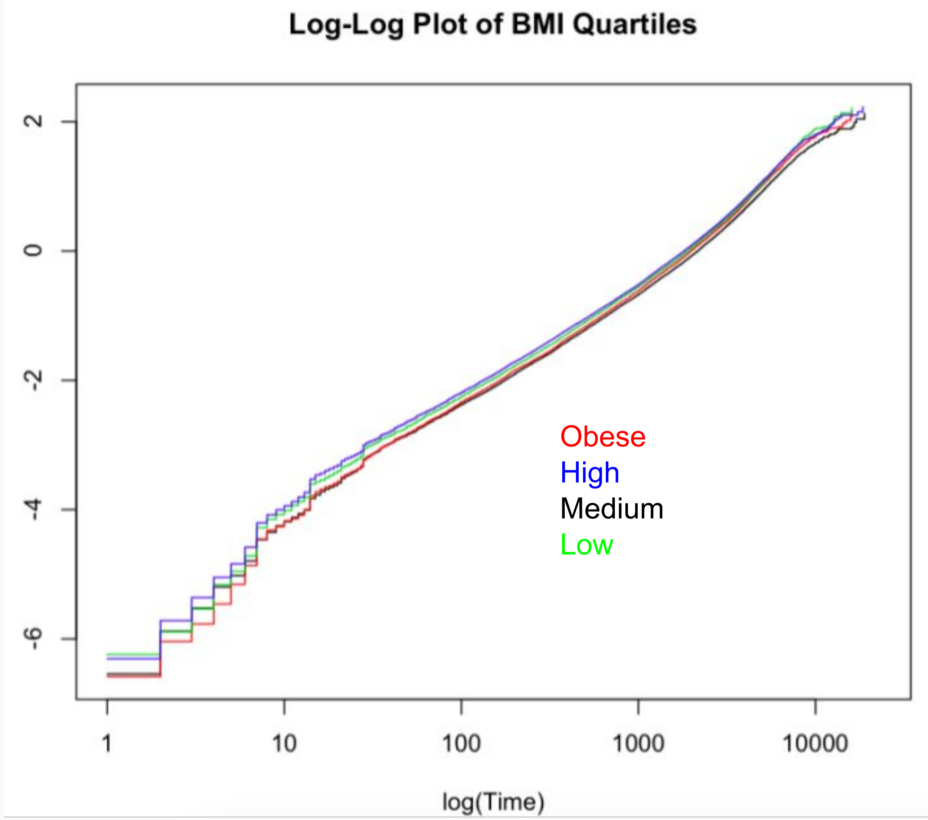
- Took advantage of the fact that we only had one observation/person
- Fit a logistic model on a one-year subset
  - Single train/test split with data between 3 and 4 years
- Used covariates from the Cox PH model
- Area Under the Curve: 77%



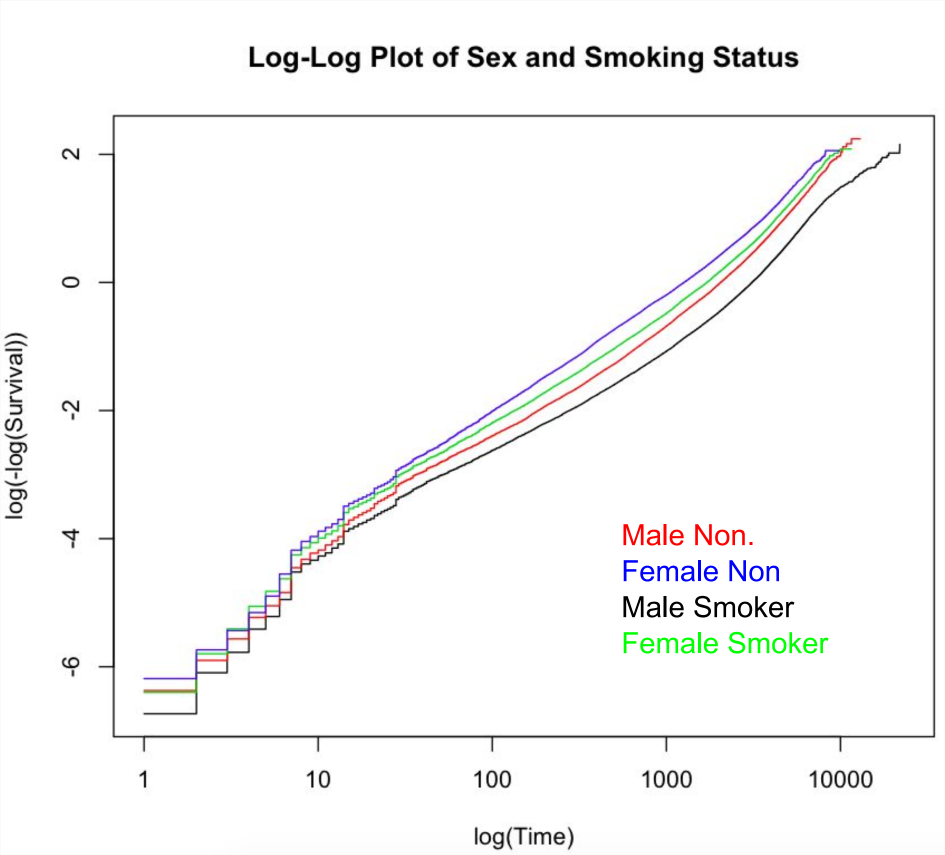
Parameterizing our Model

- Weibull model - Different assumption (Accelerated Failure Times)
    - Test via log-log plots: Linearity of log(-log(Survival Prob.) over log(Time) for each covariate-specific risk group implies assumption is valid
    - Different from PH assumption criterion which seeks parallelism between plotted lines for different risk groups for each covariate
- We also remove observations with duration > 20 years, as advised by Prof. Duncan and Nhan (thank you!)
- Dataset shrinks from 124393 to 123976 observations (- 417 observations)

Evaluating the AFT Assumption



Evaluating the AFT Assumption

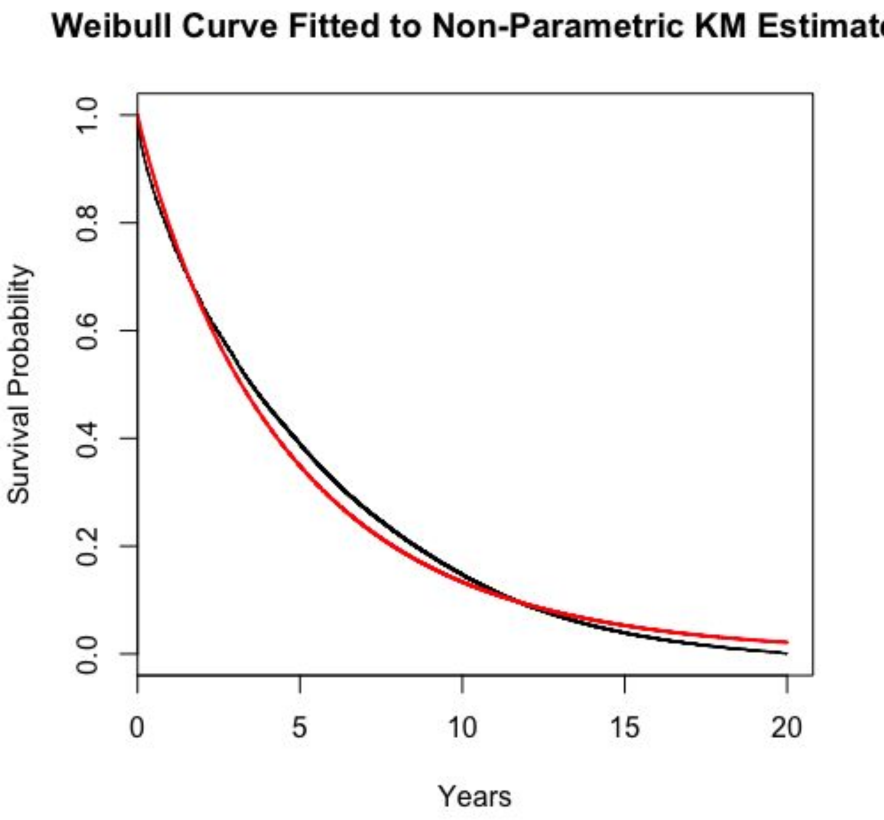


Results From Our Weibull Model

$$h(t) = \lambda^p p t^{p-1} \quad \lambda = \exp(-\sum X_i \beta_i)$$

$$p = 1.0688$$

i	X_i	β_i	Estimated HR
0	Intercept	3.409437	N/A
1	BMI	-.021305	1.021534
2	Start Age	-.024531	1.024834
3	FS	-.100441	1.105658
4	MN	.047097	0.953995
5	MS	-.002017	1.002019



Second Dataset

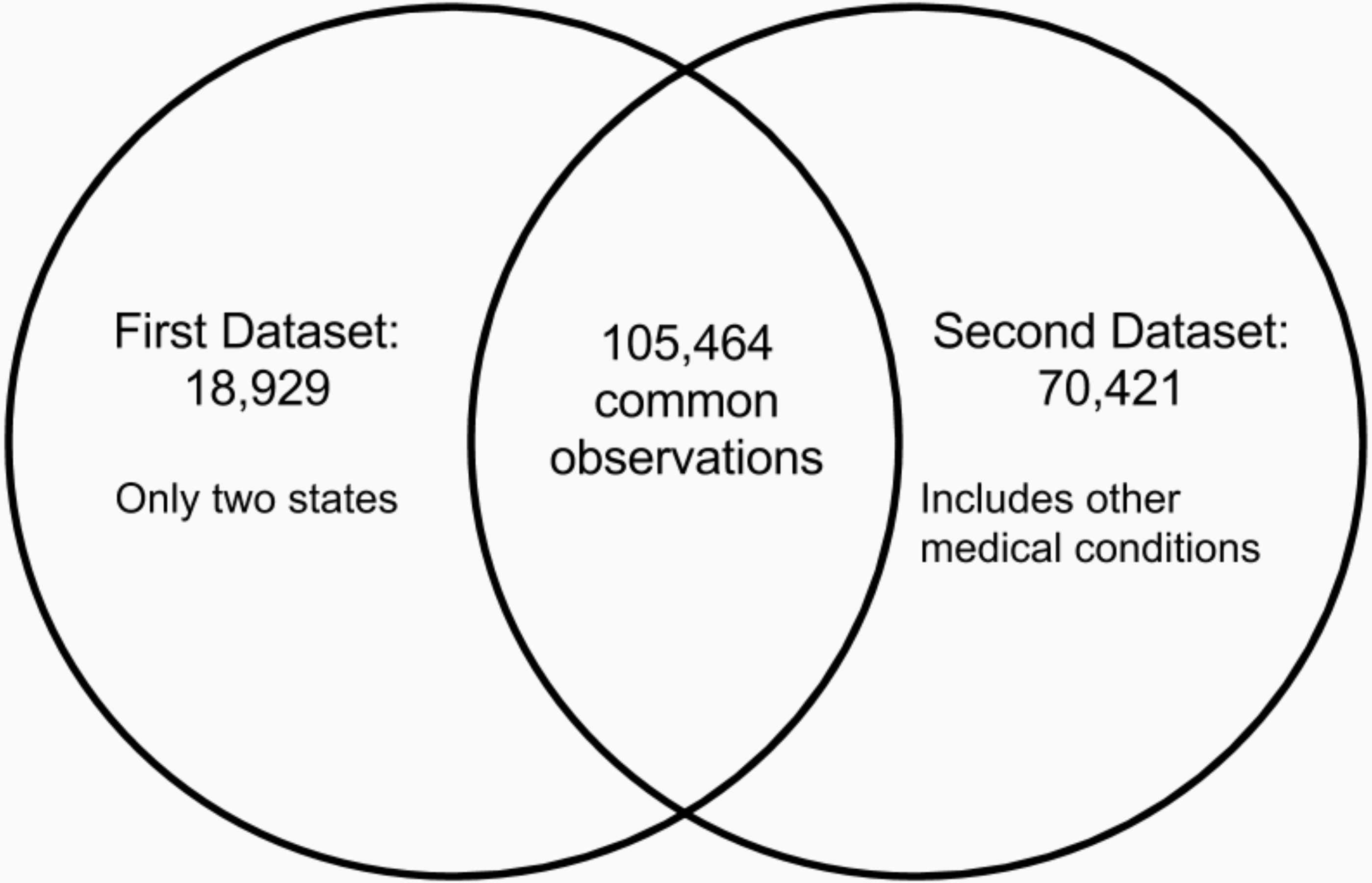
New Variables

Variable Name	Description	Support
diabetes	Binary indicator for diabetes (0: no, 1: yes)	{0,1}
hyperlipidemia	Categorical predictor for hyperlipidemia (0: none, 1: mild, 2: moderate, 3: severe)	{0,1,2,3}

Data Cleaning

Variable	Before Cleaning	After Cleaning
AGE_PRE	[-59, 102]	[8, 103]
BMInew	[0.2, 66000]	[12.1, 98.71]
DURATION (days)	[0, 21850]	(0, 21850]

- Subjects are removed for same reasons:
  - Impossible age, bmi and duration; duplicate IDs
- The size of dataset shrinks from 186,668 to 175,885:
  - 5.78% removed


First Dataset:
18,929
Only two states
105,464
common
observations
Second Dataset:
70,421
Includes other
medical conditions

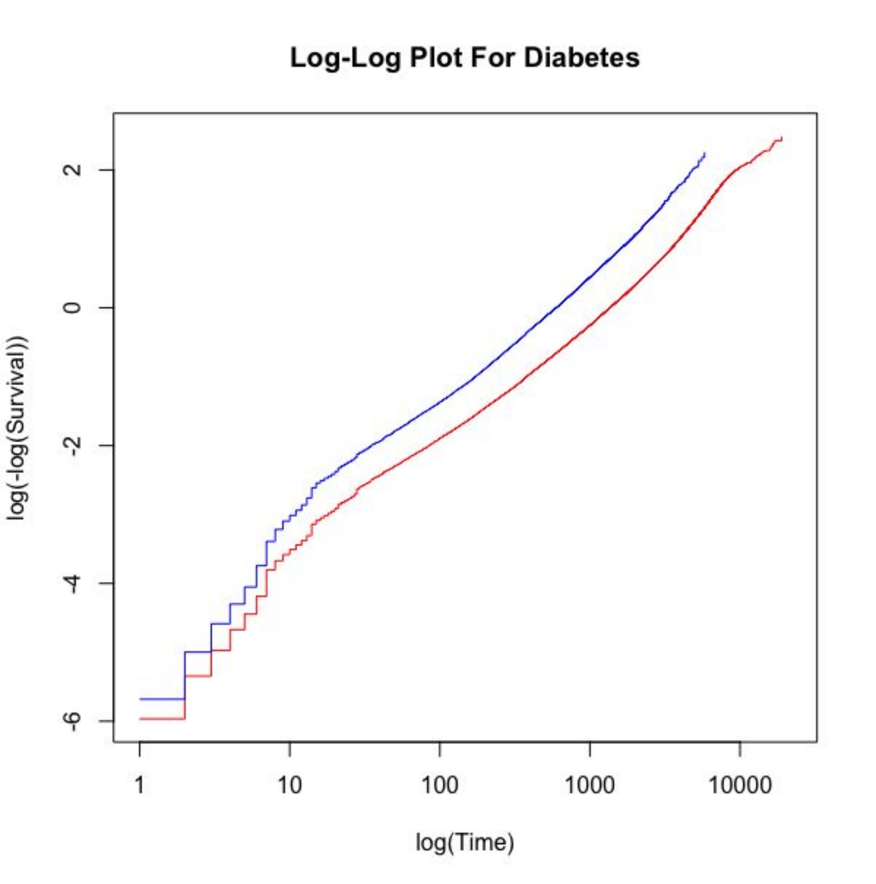
Difficulties - Mo’ Data, Mo’ Problems

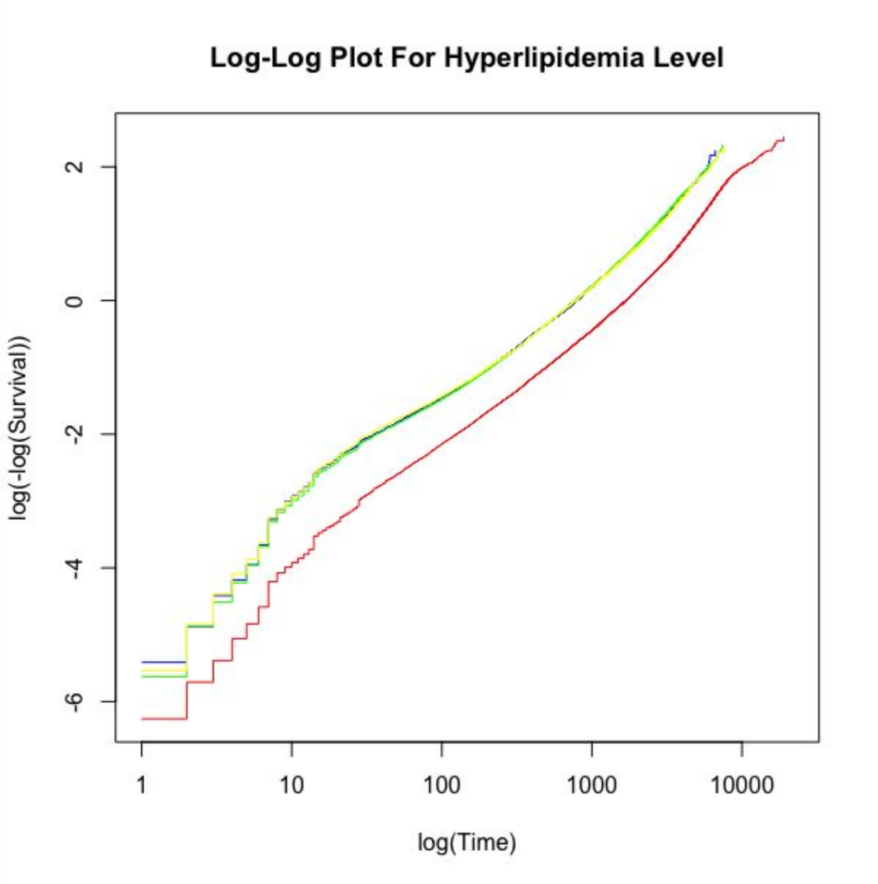
- The two datasets have significant overlap
  - However--seem to be drawn from two separate populations
- Proportional Hazards assumption not met for variables which did satisfy in old dataset - namely, SC failed in our new dataset, but was fine in the old one
- F-tests for variance between old and new datasets:
  - Ratio of variances for BMI: 0.9575417
  - Ratio of variances for duration: 1.29357

Fitting a Model

- We specifically wanted to fit a model using the new predictors for hyperlipidemia and diabetes
- PH assumption met for both hyperlipidemia and diabetes, but not SC
- Natural decision: Build Cox model stratifying on SC, using hyperlipidemia and diabetes as predictors
- 4 levels of SC - 4 differing baseline hazard rates

PH Assumption Verification For New Predictors





Our Second Stratified Model's Results

$$h(t) = h_{sc}(t) \cdot \exp(0.622 \cdot hyper1 + 0.617 \cdot hyper2 + 0.598 \cdot hyper3 + 0.464 \cdot diabete)$$

$$sc = FN, FS, MN, MS$$

Covariate	Estimated Hazard Rate	95% Confidence Interval
Hyperlipidemia (level 1)	1.862	[1.828, 1.897]
Hyperlipidemia (level 2)	1.853	[1.829, 1.878]
Hyperlipidemia (level 3)	1.814	[1.787, 1.843]
Diabetic	1.590	[1.562, 1.619]

Baseline Hazard Rates for Different Sex-Cigar Strata

Time = 1 year

Sex-Cigar	Rate
FN	0.3666378
FS	0.4207363
MN	0.3512549
MS	0.3915465

Time = 3 years

Sex-Cigar	Rate
FN	0.9009510
FS	0.9957537
MN	0.8334356
MS	0.9027204

Time = 5 years

Sex-Cigar	Rate
FN	1.431580
FS	1.529047
MN	1.308066
MS	1.377740

Future Directions

- The data we used for this project is a subset of a dataset to be used in PSTAT 296 for the upcoming year
  - Recommend carefully checking any new predictors if a new subset is taken
- Change event of interest to diabetes, hyperlipidemia, etc.
  - Use hypertension as predictor
- Include multiple observations per person if possible
  - Better evaluate time-dependency
- Possibly using data from sites such as Kaggle, Reddit, UCI
  - Kaggle hosts datasets from companies and government agencies

Acknowledgements

- 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

Q & A

Thank you!