Risk Modeling for Railroad Incidents: Variable Screening, Model Type Selection, Functional Form, and Model Development

October 24, 2024

2024 FCSM Research and Policy Conference, Hyattsville, Maryland

Young-Jun Kweon, Mathematical Statistician, Bureau of Transportation Statistics, USDOT
Jianqiang (Tony) Ye, Operations Research Analyst, Federal Railroad Administration, USDOT
Ruby Li, Operations Research Analyst, Federal Railroad Administration, USDOT

Disclaimer

The opinions, findings, and conclusions expressed in this presentation are those of the authors and not necessarily those of the United States Department of Transportation, the Federal Railroad Administration, or the Bureau of Transportation Statistics. The United States Government assumes no liability for its contents or use thereof.

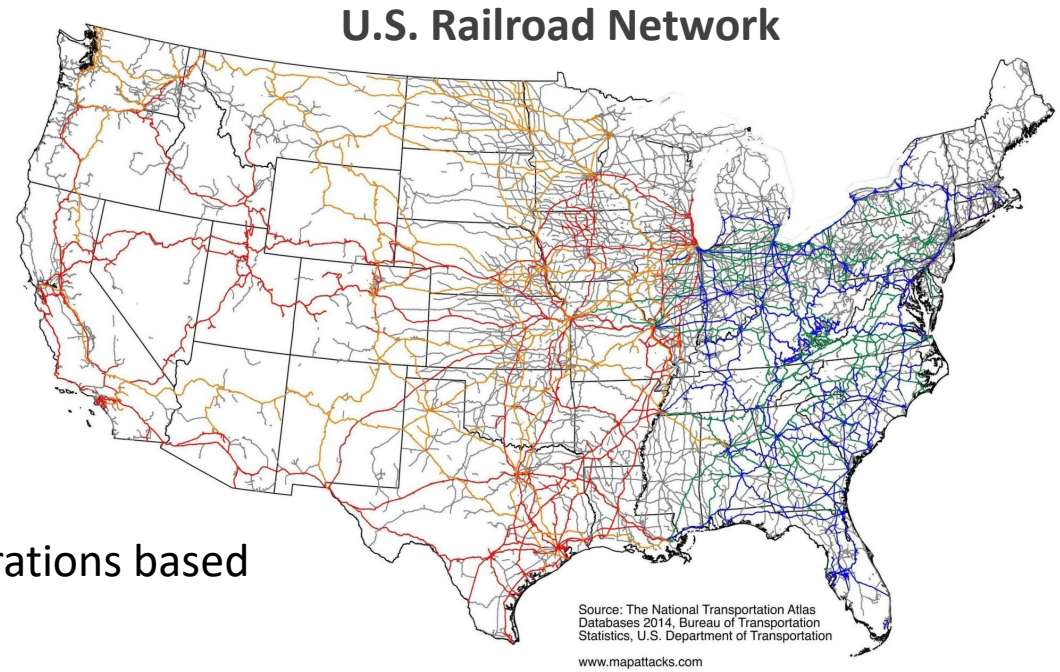
Acknowledgement

Thanks go to:

- Andrew LaBounty (Transportation Analyst GIS), Brian Shiner (Transportation Analyst), Patrick Johnson (Data Engineer), and Raquel Wright (Geospatial Information Officer) for providing technical support and consultation in data collection and preparation
- Emily Grenzke (Staff Director) for providing continuous support and guidance throughout the project
- Gary Fairbanks (Staff Director) and Doug Yates (Deputy Staff Director) for subject matter expertise related to locomotive and equipment risks.

FRA Risk Modeling Objectives

- FRA promotes and regulates railroad safety.
- There are six technical disciplines in FRA's Safety Inspection Program that cover different aspects of railroad safety compliance and enforcement.



- For each discipline, FRA has developed risk modeling:
  - ❑ To provide inspectors and specialists risk-based information that supports development of Focused Inspection Plans
  - ❑ To identify current and future risks of railroad assets and operations based on the best available data
  - ❑ To focus our efforts and resources on those areas likely having higher risk of incidents, casualties, or damages by complementing field experience with data analysis
  - ❑ To perform risk assessments on identified hot spots

MP&E Risk Modeling

- The Motive Power and Equipment (MP&E) Division promotes an understanding of and compliance with Federal standards to inspect locomotives, passenger and freight cars, and safety appliances such as air brakes.
- MP&E risk modeling informs efforts to allocate resources likely to have a higher risk of incidents caused by locomotive and equipment failure or malfunction.
- The Norfolk Southern train derailment in East Palestine Ohio in 2023 is an example of equipment failure. According to the NTSB final report, a rail car's defective wheel bearing caused the derailment and subsequent hazardous material release.
- FRA developed its first generation of MP&E risk model in early 2022 and used Tableau as the visualization tool.
- Today we are discussing the data and methodologies we have been using for the second generation of MP&E risk modeling.

7-Step Process for Developing Predictive Risk Model

STEP 1	Examine Data	Develop insight on variables
STEP 2	Treat Missing	Decide how to treat missing values
STEP 3	Select Model Type	Select an appropriate model type
STEP 4	Screen Variables	Screen out not useful variables
STEP 5	Select Specific Model	Select specific model(s) within the selected model type
STEP 6	Select Variables	Select variables and functional forms
STEP 7	Develop Final Model	Estimate alternative models and select the best one

STEP 1: Examine Data ①

MP&E Data: 161 variables from 8 data sources

Develop insight on variables

Data Source	# of Variables
NARN	77
AIRS	25
Waybill Sample	14
Census Bureau	14
Form 96 Inspections	12
Form 54 Rail Eq. Incidents	10
Form 97 Accountables	4
Form 55a Injuries/Illnesses	2
Derived	2
Total	161

The Forms listed here are required under 49 CFR Part 225:

- Inspections are completed by FRA staff or state partner inspection programs.
- Reportable incidents meet the total damage cost threshold, currently $12,000, as recorded by the railroads.
- Accountable incidents are initially reported by the railroads but do not meet the damage threshold.
- Reportable injuries and illnesses similarly meet threshold criteria for reporting to the FRA.

STEP 1: Examine Data ②

Look at 4 attributes

- Variable type
- Number of unique values
- Number of missing cases
- Range of values

MP&E Data

Name	Type	# of Unique Values	# of Missing Cases	Value Range
PK	numeric	3316	0	[1 - 3347]
AIRSCode	character	1126	200	[MP&E101-NJ-BRW-31540 - MP&E813-WA-YCR-12660]
TerritoryCode	character	91	0	[MP&E101 - MP&E813]
District	numeric	8	0	[1 - 8]
InspectorPayrollId	numeric	84	200	[10367 - 986]
InspectorName	character	83	270	[Ackerman, Justin - Wozniak, Thomas M.]
OrganizationCode	character	660	0	[AA - ZWSX]
OrganizationName	character	659	0	[1003 OPERATIONS (XLLT) - Zanesville & Western Scenic Railroad]
OrganizationTypeCode	character	2	0	[C - R]
StateFIP	numeric	50	0	[1 - 56]
StateAbbreviation	character	50	0	[AK - WY]
CountyFIP	character	198	0	[C001 - C840]
CountyName	character	1003	0	[ADAMS - YUMA]
CityFIP	numeric	967	0	[8 - 962]
CityName	character	1834	0	[ABERDEEN - ZANESVILLE]
FacilityName	character	2854	0	[(DCTA) MAINTENANCE FACILITY - ZWSX - ZANESVILLE AND WESTERN SCENIC RAILROAD]
Latitude	numeric	3183	0	[25.85363 - 64.84839]
Longitude	numeric	3180	0	[-100.00636 - -99.92862]
AverageDailyTrains	numeric	65	0	[0 - 43]
AverageDailyCars	numeric	281	0	[0 - 469]
AverageDailyLocomotives	numeric	75	0	[0 - 228]
NumberCarShops	numeric	12	0	[0 - 111]
NumberLocomotiveShops	numeric	7	0	[0 - 111]
LatLong	character	3199	0	[POINT (-100.006362 37.750554) - POINT (-99.928623 47.769654)]
f54_UniqueIncidents	numeric	31	2053	[1 - 72]
f54_MinMetersFromAIRS	numeric	1221	2053	[13 - 142705]

STEP 2: Treat Missing ①

3 options for treating missing values

Decide how to treat missing values and treat them accordingly

- Edit by rules
- Impute by imputation model
- Do nothing

How to make a choice?

- Based on insight from STEP 1 & input from and consultation with subject matter experts (SMEs)

Example: MP&E Data

- f54_UniqueIncidents records the number of incidents related to MP&E safety discipline.
- There are 2,053 missing cases out of 3,316.
- Missing means there was no incident according to SMEs.
- **Edit (Replace missing by zero)**

STEP 2: Treat Missing ②

MP&E Data

Missing Treatment	# of Variables
No missing	*63*
Do nothing	72
Edit by Missing = 0, 9, or "X"	21
Impute by model	5
Total	161

STEP 3: Select Model Type ①

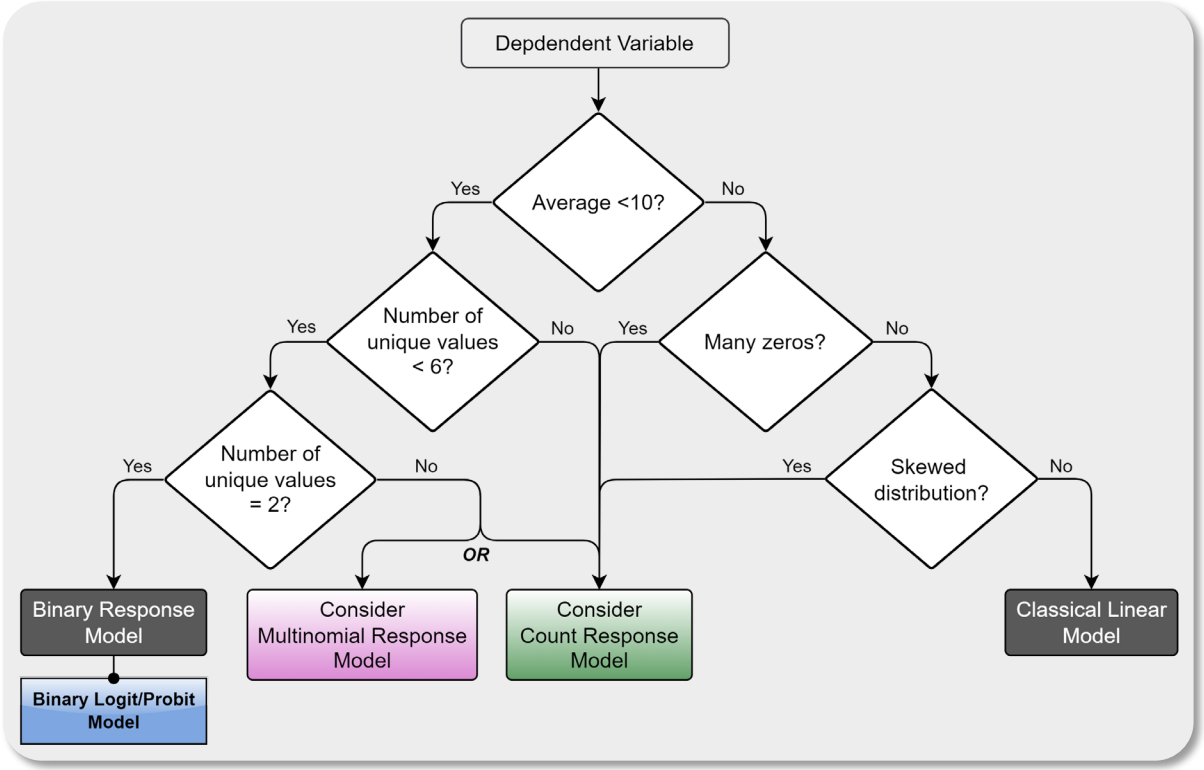
Regression Analysis

Select an appropriate model type

- Various model types exist
- Characteristics of the dependent variable dictates an appropriate model type

Decision Chart

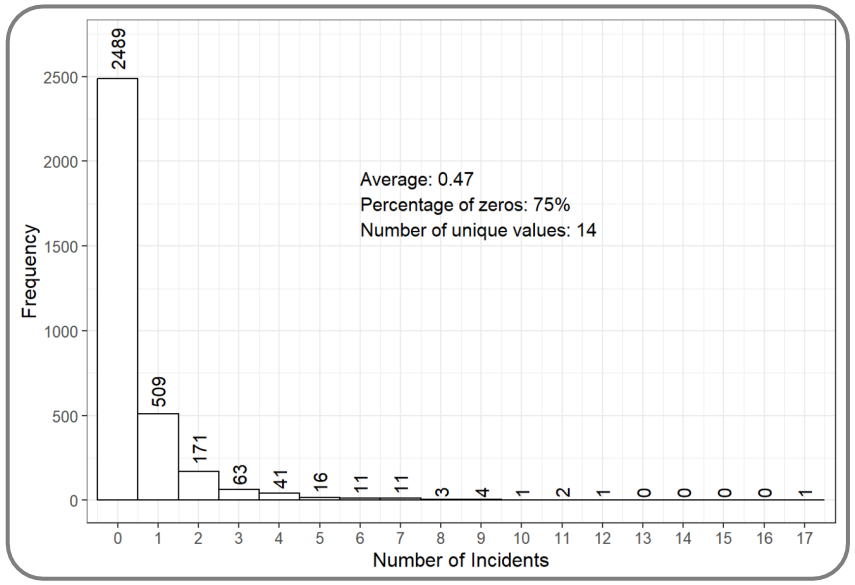
- To guide selection of appropriate model type

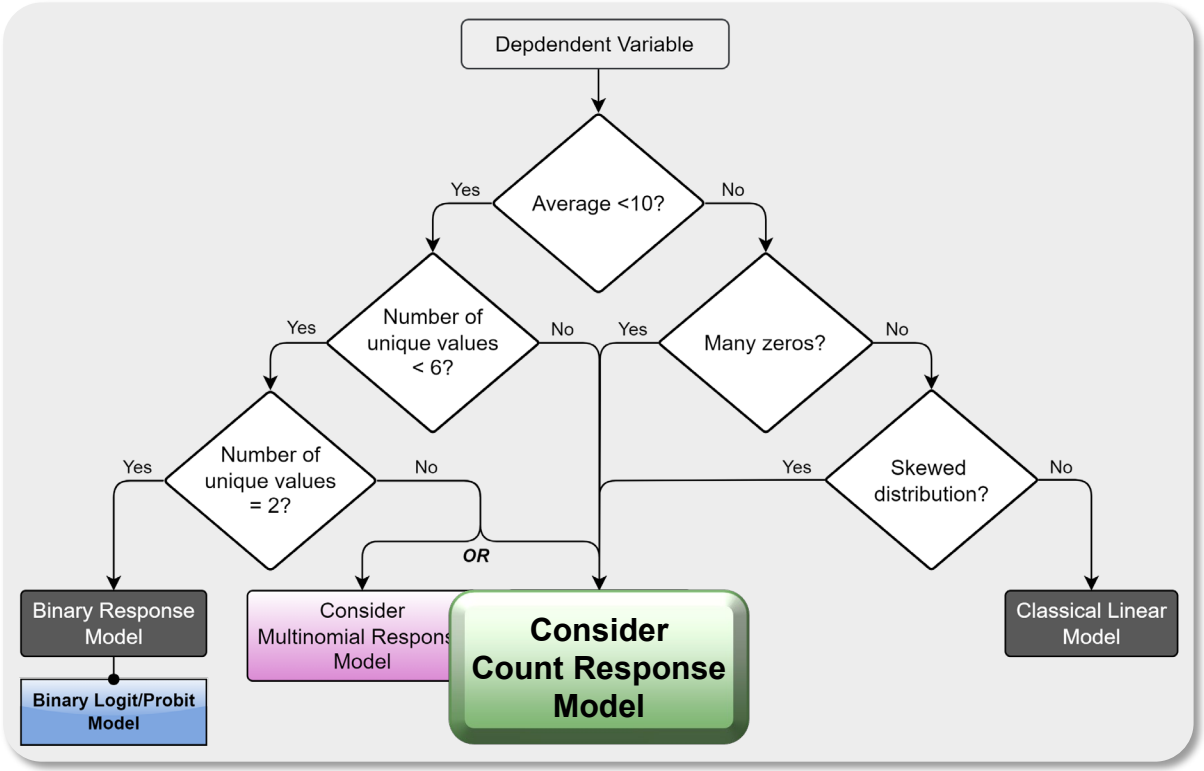


STEP 3: Select Model Type ②

MP&E Data

- Dep. Var: Number of MP&E Incidents in 5 year (2018-2023)





STEP 4: Screen Variables

3-Stage Screenings

Screen out not useful variables

- First Screening (based on variable definition, variable type, and # of unique values)
- Second Screening (based on variable type, # of missing values, logical consideration, and subject matter expert review)
- Third Screening (based on variable definition, logical consideration, and statistical test)

MP&E Data

Screening	# of Removed Variables
First Screening	77
Second Screening	35
Third Screening	3
Total	115

STEP 5: Select Specific Model ①

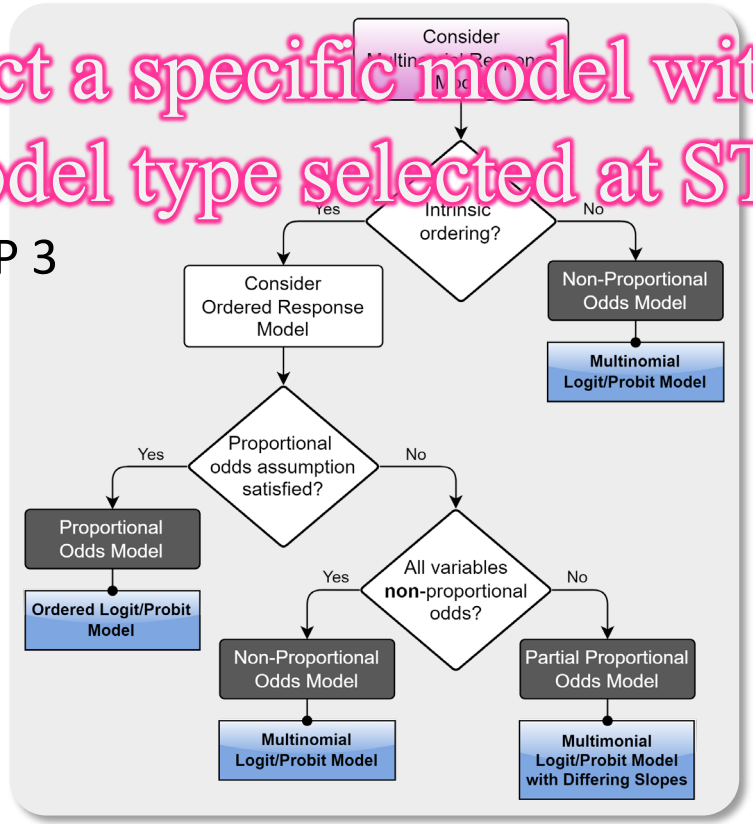
STEP 5

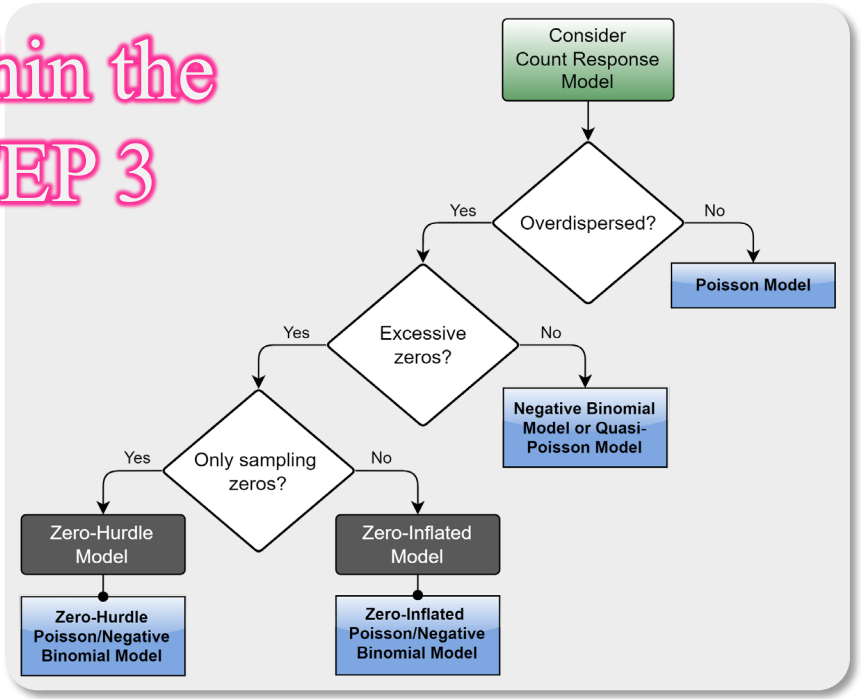
- is skipped or
- leads to another decision chart depending on the decision at STEP 3

Select a specific model within the model type selected at STEP 3

Two Decision Charts

- Multinomial Response Model
- Count Response Model

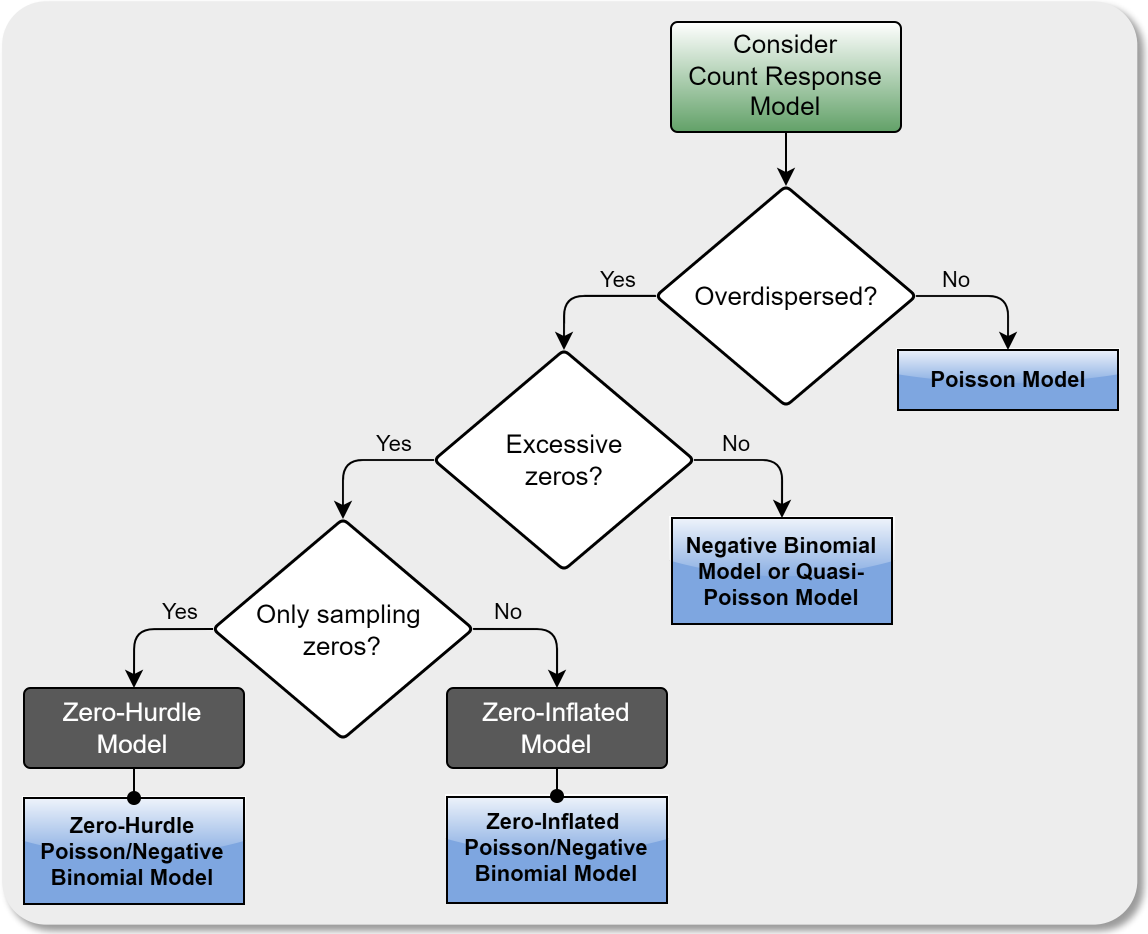




STEP 5: Select Specific Model ②

MP&E Data

- Overdispersed?
  - Confirmed by Lagrange Multiplier (LM) test and 2 NB dispersion parameter tests
- Excessive zeros?
  - Determined by percentage of zeros (75%)
- Only structural zeros?
  - Undecided
- Selection:
  - Zero-Hurdle Models (ZHP, ZHNB)
  - Zero-Inflated Models (ZIP, ZINB)



STEP 6: Select Variables ①

3 Tasks for Selecting Variables

- Create new variables by combining existing variables
- Devise alternative functional forms suggested by using EDA and SAM
- Select predictor variables and their forms

Select variables and functional forms in the model selected at STEP 5

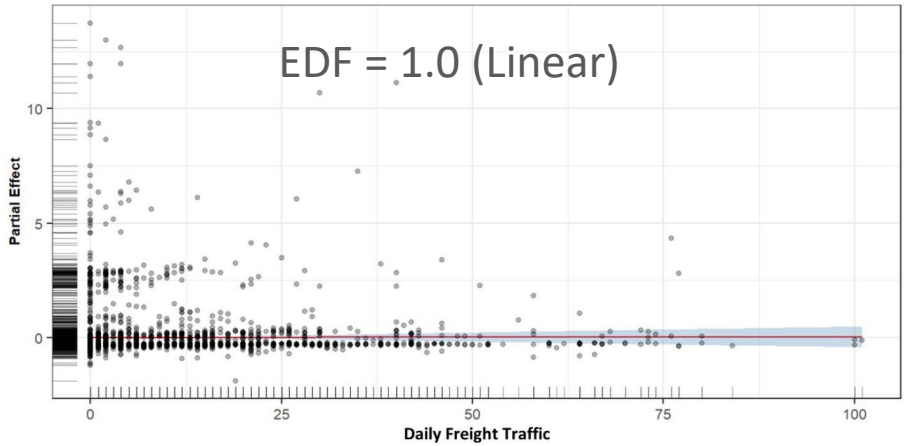
STEP 6: Select Variables ②

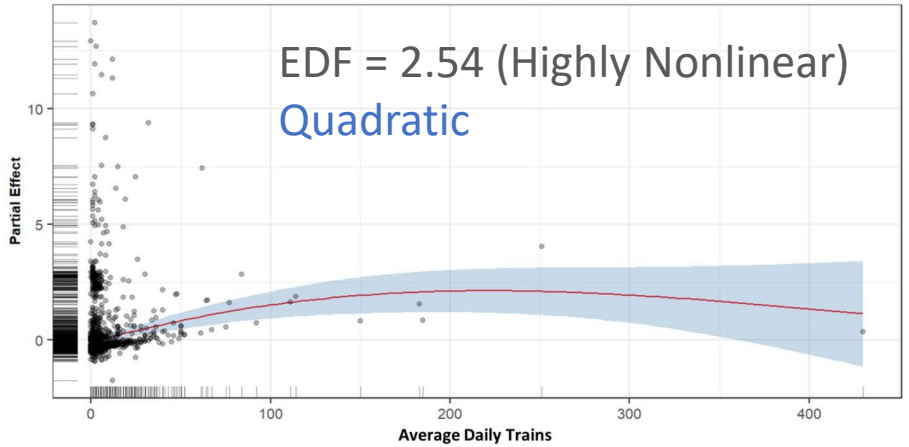
MP&E Data

- Effective Degree of Freedom (EDF)

EDF	Implication
EDF = 1	Linear
1 < EDF < 2	Weakly Nonlinear
EDF ≧ 2	Highly Nonlinear

- Generalized Additive Model (GAM)





STEP 7: Develop Final Model ①

3 Tasks for Developing Final Model

- Specify candidate models
- Estimate candidate models
- Select the best model

Estimate all candidate models and select the best one

MP&E Data: 4 Candidate Models

Statistics	ZIP	ZINB	ZHP	ZHNB
Num. of Observations	3,323			
Log-Likelihood	-2,722	-2,654	-2,784	-2,787
AIC	5,495	5,339	5,624	5,515
BIC	5,648	5,430	5,795	5,680
AICc	5,495	5,339	5,624	5,516
Vuong Statistic (p-value): ZIP vs. ZINB	-3.067 (0.0011)			
Vuong Statistic (p-value): ZHP vs. ZHNB	-3.323 (0.0004)			
Vuong Statistic (p-value): ZINB vs. ZHNB	4.635 (0.000001)			

STEP 7: Develop Final Model ②

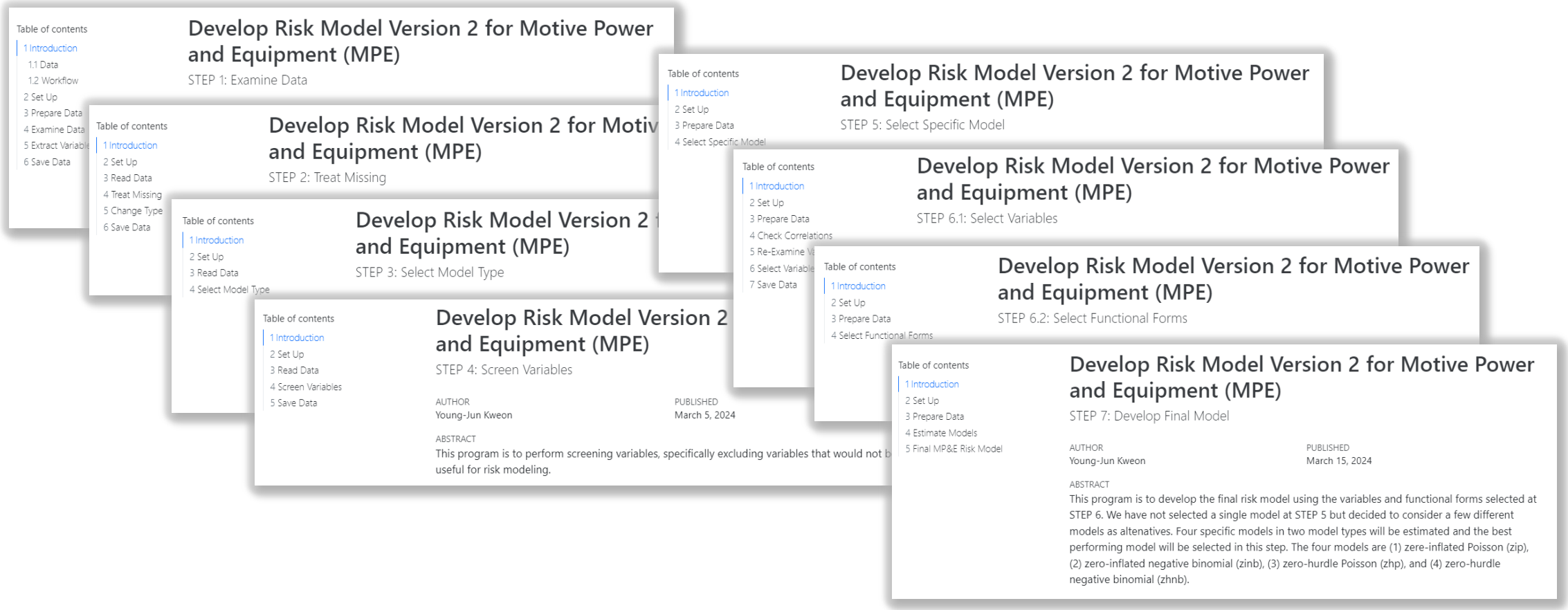
MP&E Data: Final Model

- ZINB (Zero-Inflated Negative Binomial)

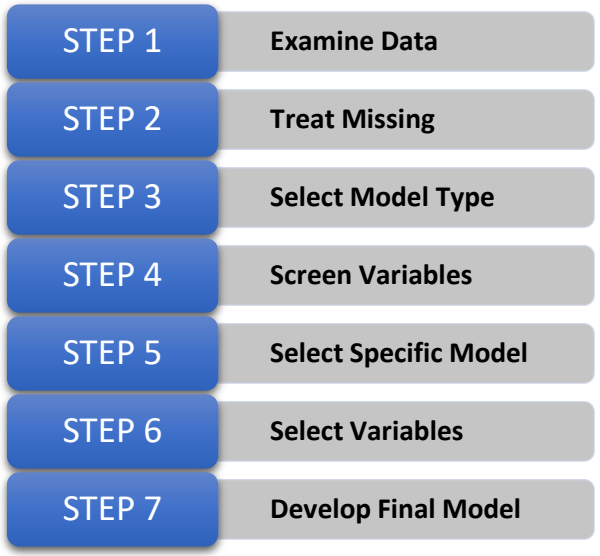
Variables	ZINB		
	Coeff.	Std. Err.	p-value
Count Component Model (Negative Binomial)			
(Intercept)	-1.0955	0.0895	0.0000
Daily Freight Traffic	0.0124	0.0028	0.0000
HazMat Cars	0.0369	0.0032	0.0000
Accountable Incidents	0.0475	0.0072	0.0000
Average Daily Trains	0.0125	0.0044	0.0045
Locomotive Shops	0.4044	0.0736	0.0000
Indicator (District: 2 or 8)	-0.2892	0.0867	0.0009
Indicator (Signal Type: ACS)	0.7652	0.3813	0.0448
Indicator (Signal Type: MAN)	-0.3738	0.0954	0.0001
Indicator (Signal Type: TWC)	-0.4543	0.1330	0.0006
Indicator (Track Class: 4)	-0.1990	0.0902	0.0274
Dispersion (k)	1.2079	0.8987	0.0770
Binary Component Model (Logit)			
(Intercept)	-0.6812	0.2893	0.0185
Average Daily Cars	-0.0177	0.0065	0.0064
Num. of Observations	3,323		
Log-Likelihood	-2,654		

Documentation

8 html documents were created by Quarto



Takeaways



- 7-step process enhances consistency and efficiency in developing predictive risk models across six inspection disciplines.
  - 4 attributes (variable type, number of unique values, number of missing cases, range of values) examined at STEP 1 facilitate decision on missing treatment (STEP 2) and variable screening (STEP 4)
  - Selecting multiple models might be inevitable at STEP 5, especially count data with overdispersion.
  - GAM at STEP 6 is useful in suggesting functional forms yet cumbersome when there are many numeric variables.
- Good documentation is strongly desired for accountability and reproducibility.

Contact Us

Federal Railroad Administration
1200 New Jersey Avenue, SE
Washington, DC 20590

Connect with us at USDOTFRA

Young-Jun Kweon
Email: young-jun.Kweon@dot.gov

Jianqiang (Tony) Ye
Email: jianqiang.ye@dot.gov

Ruby Li
Email: ruby.li@dot.gov

U.S. Department of Transportation
Federal Railroad Administration