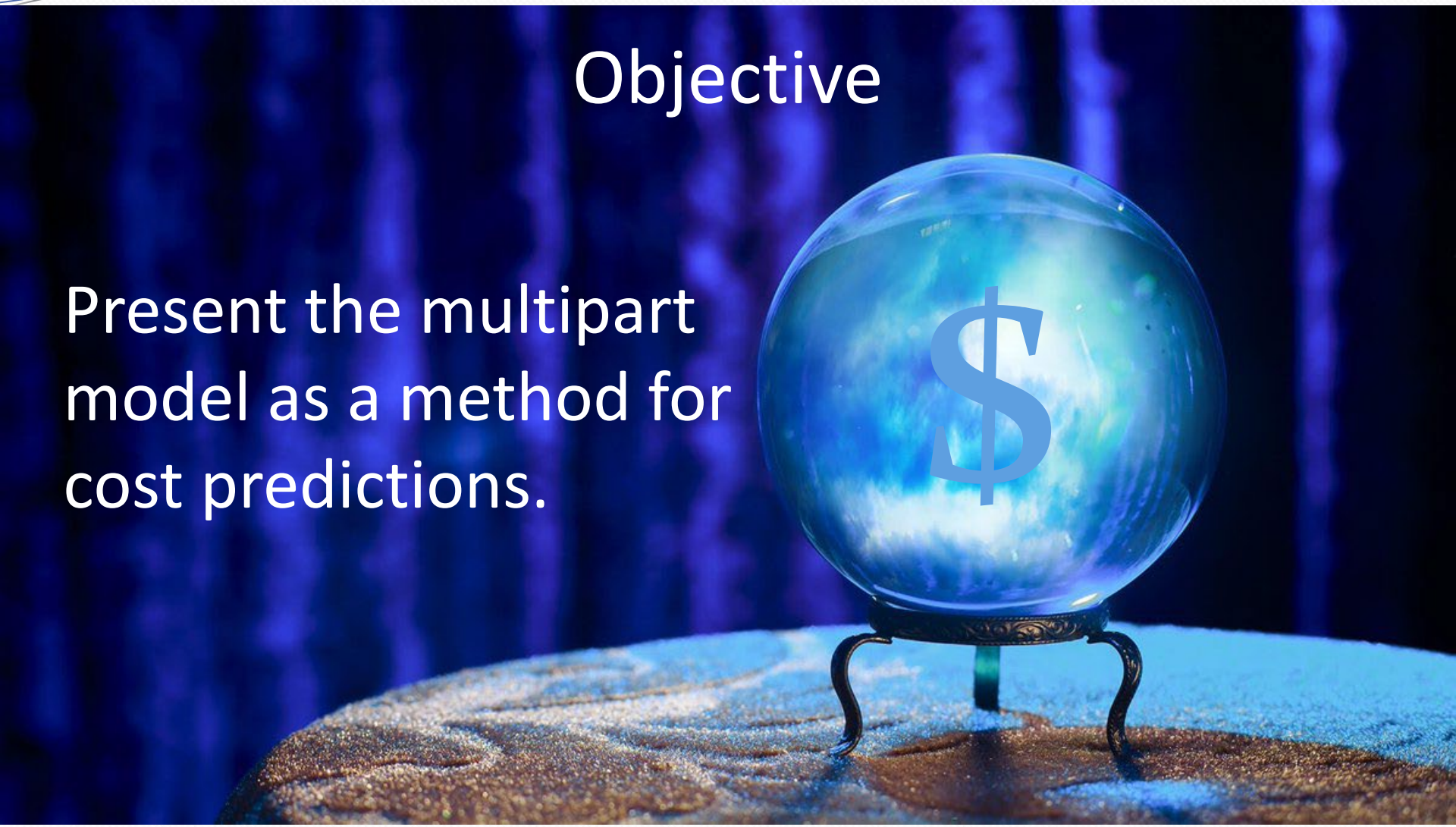
March 15, 2023 Health Economics Resource Center Presentation

Multipart models of continuous outcomes

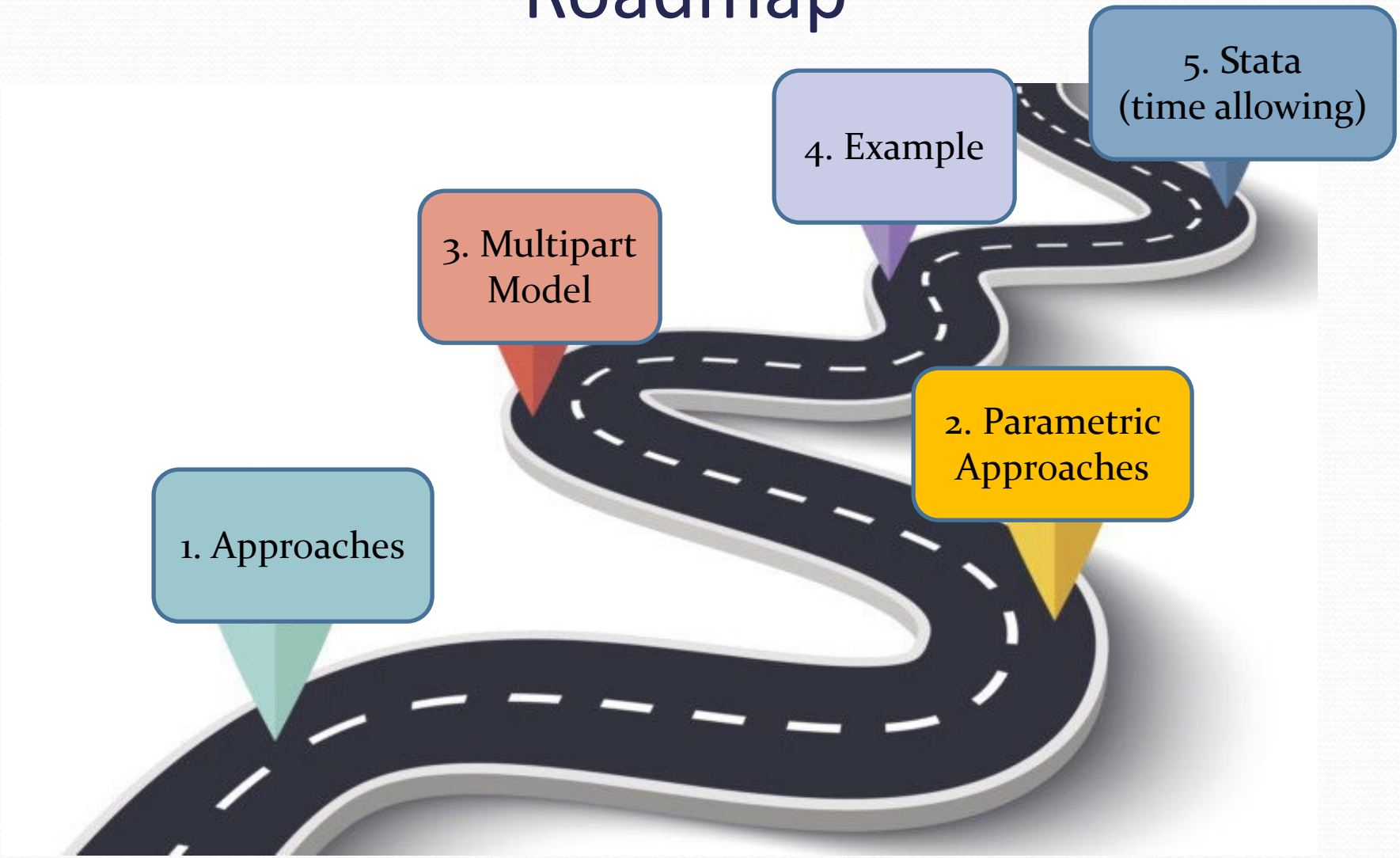
Peter Veazie, PhD
GECDAC
Canandaigua VAMC
peter.veazie@va.gov
peter_veazie@urmc.rochester.edu

Objective

Present the multipart model as a method for cost predictions.

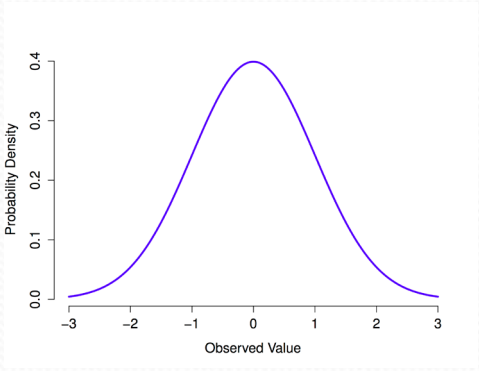


Roadmap

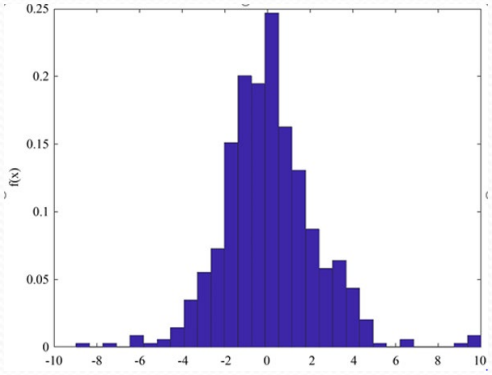

5. Stata
(time allowing)
4. Example
3. Multipart
Model
2. Parametric
Approaches
1. Approaches

Approaches

- **Parametric** estimation (the subject of this presentation)
  - e.g., define functional form of conditional parameters (e.g. regression and skedastic functions).
  - e.g., define functional form of distributions

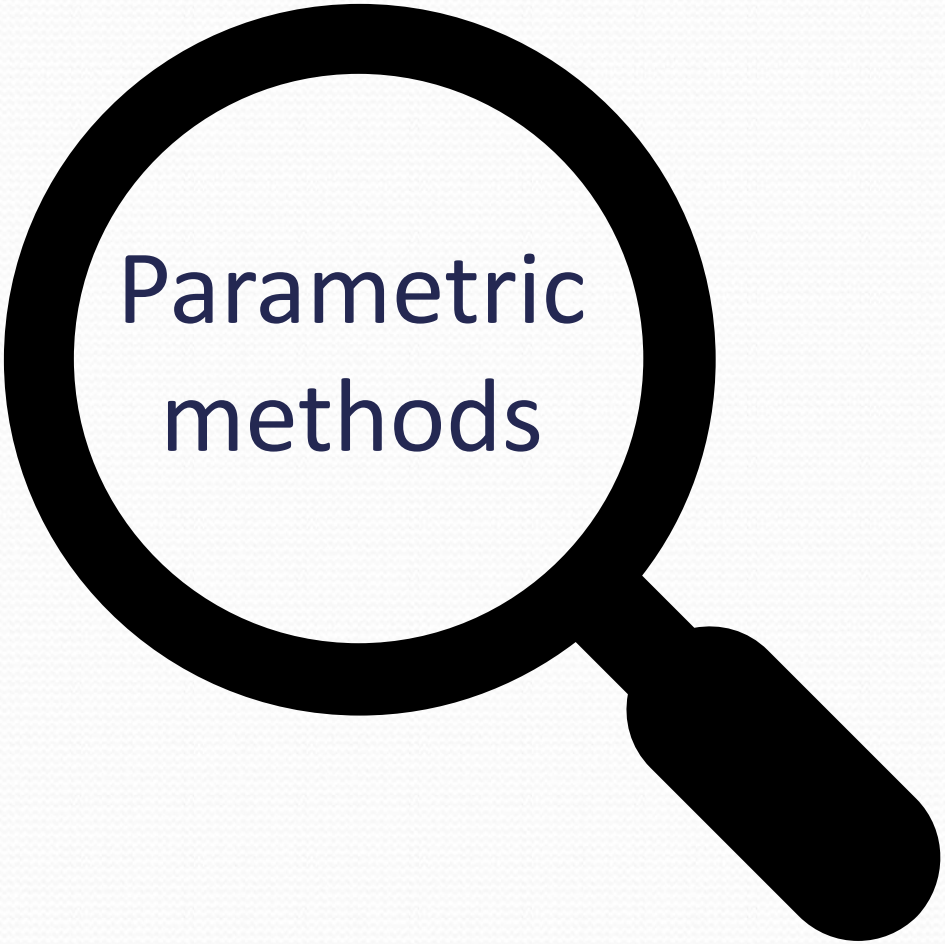


- **Nonparametric** approaches
  - Kernel regression: weighted averages for each observation
  - Series regression: using linear combinations of basis functions



- **Semiparametric** estimation
  - Partially linear models: $y = \beta' X + g(Z) + \varepsilon$
  - Index models: $y = g(\beta' X) + \varepsilon$

- **Machine learning** approaches
  - Classification and regression trees
  - Random forests
  - Generalized boosting methods


Parametric
methods

Transformations

- Express transformed dependent variable (e.g., cost) as linear function of predictors:

$$g(Y) = \beta' \mathbf{X} + \varepsilon$$

$$Y = g^{-1}(\beta' \mathbf{X} + \varepsilon)$$

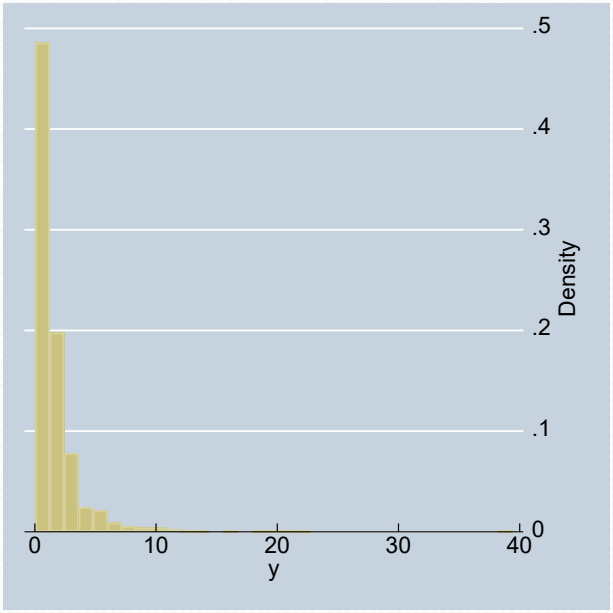
Prediction based on

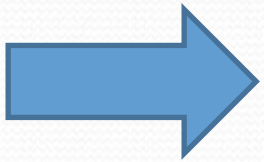
$$\hat{E}(Y \mid \mathbf{X}) = \int g^{-1}(\hat{\beta}'\mathbf{X} + \varepsilon) \cdot d\hat{F}(\varepsilon)$$

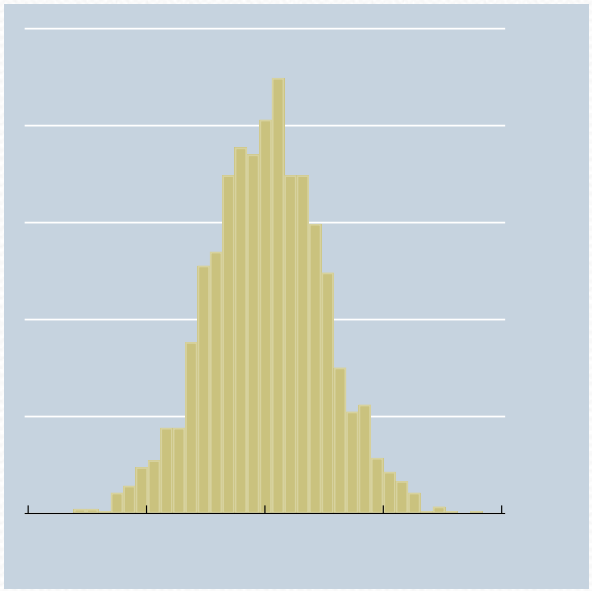
- Specify a distribution of errors
- Or, integrate out error using estimated residuals

Transformation examples

Example: logarithm transformation







Y

$\mathrm{Ln}(Y)$

Transformed variable as a linear combination of predictors

$$\ln(Y) = \beta'\mathbf{X} + \varepsilon$$

$$\hat{E}(\ln(Y) \mid \mathbf{X}) = \hat{\beta}'\mathbf{X}$$

"Congress does not allocate log-dollars"
Will Manning

Retransform to predict *Y* rather than ln(*Y*)

$$Y = e^{\beta' \mathbf{X} + \varepsilon}$$

$$\hat{E}(Y \mid \mathbf{X}) = e^{\hat{\beta}' \mathbf{X}} \int e^{\varepsilon} \cdot d\hat{F}(\varepsilon)$$

If assume *F* is $N(0,\sigma^2)$:

$$\hat{E}(Y \mid \mathbf{X}) = e^{\hat{\beta}' \mathbf{X} + 0.5 \cdot \hat{\sigma}_{\varepsilon}^2}$$

If not assume *F* has a specific distribution:

$$\hat{E}(Y \mid \mathbf{X}) = e^{\hat{\beta}' \mathbf{X}} \cdot \underbrace{\frac{1}{n} \sum_{i} e^{\hat{\varepsilon}_i}}_{\text{Duan smearing estimator}}$$

Example: square root transformation

$$\sqrt{Y} = \beta'\mathbf{X} + \varepsilon$$

$$Y = (\beta'\mathbf{X} + \varepsilon)^2$$

$$\hat{E}(Y \mid \mathbf{X}) = \left(\hat{\beta}'\mathbf{X}\right)^2 + \hat{\sigma}_\varepsilon^2$$

Nonlinear regression functions

Directly model conditional expectation as a function of the linear predictor:

$$E(Y \mid \mathbf{X}) = h(\beta' \mathbf{X})$$

- Estimate using nonlinear least squares based on

$$Y = h(\beta' \mathbf{X}) + \varepsilon$$

- Estimate using MLE/QMLE (GLM models)

$$h^{-1}\left(E(Y \mid \mathbf{X})\right) = \beta' \mathbf{X}$$

h^{-1} is the link function

- Common link functions for costs
  - Natural logarithm
  - Square root
- Common distributions for costs
  - Gamma
  - Generalized gamma
  - Sing-Maddala
  - Log-logistic
  - Beta prime
  - Tweedie

Multipart model

Partition the dependent variable into regions

Expand the conditional expectation across the regions:

$$E(Y \mid \mathbf{X}) = \sum_{r \in R} E(Y \mid \mathbf{X}, r) \cdot P(r \mid \mathbf{X})$$

Estimate individual parts

$$\hat{E}(Y \mid \mathbf{X}) = \sum_{r \in R} \underbrace{\hat{E}(Y \mid \mathbf{X}, r)}_{A} \cdot \underbrace{\hat{P}(r \mid \mathbf{X})}_{B}$$

Estimate each part in *A* using any of the preceding methods for each region

Estimate each part in *B* using

- Ordered categorical variable model (e.g. ordered Logit)
- Multinomial model

Example—GEC population

Objective: To compare performance of cost models on the Geriatrics and Extended Care (GEC) population of Veterans

Setting: Veterans in the United States Department of Veterans Affairs (VA) not residing long-term in nursing homes (noninstitutional) and without end-stage-renal-disease (ESRD).

Data Source: Fiscal year 2017 costs and diagnoses from VA inpatient, outpatient, drug, and prosthetics Managerial Cost Accounting data, VA purchased care claims, socio-demographic enrollment and vital status data, Medicare enrollment, and diagnoses from inpatient/outpatient and Carrier claims data.

Study Design: Models were estimated on noninstitutional Veterans and the subpopulation receiving GEC services.

Risk-adjusters included age, indicators of being white, male, married, having insurance, being on a VA chronic illness registry, 8 VA priority status groups, 24 drug classes, 47 mental health conditions, 84 Hierarchical Clinical Conditions, and the JEN Frailty Index. Models were compared by mean residuals, R-square, and root mean squared errors (RMSE).

Models

Model	Description	Prediction equation	Ancillary equations/specifications
SH	Square-root Transformed costs with homoscedastic errors	$\hat{E}(\text{Cost} \mid X) = (X'\hat{\beta})^2 + \hat{V}(\varepsilon)$	$\hat{V}(\varepsilon) = MSE(\hat{\varepsilon})$
SL	Square-root Transformed costs with log-transformed heteroskedastic errors	$\hat{E}(\text{Cost} \mid X) = (X'\hat{\beta})^2 + \hat{V}(\varepsilon \mid X)$	$\ln(\hat{\varepsilon}) = X\theta + \eta_\varepsilon$ $\hat{V}(\varepsilon \mid X) = e^{X\cdot\hat{\theta}+0.5\cdot MSE(\hat{\eta}_\varepsilon)}$
SD	Square-root Transformed costs with Duan heteroskedastic errors	$\hat{E}(\text{Cost} \mid X) = (X'\hat{\beta})^2 + \hat{V}(\varepsilon \mid X)$	$\ln(\hat{\varepsilon}) = X\theta + \eta_\varepsilon$ $\hat{V}(\varepsilon \mid X) = e^{X\cdot\hat{\theta}} \cdot \frac{1}{N}\sum_{i=1}^{N} e^{\eta_i}$
SS	Square-root Transformed costs with square-root heteroskedastic errors	$\hat{E}(\text{Cost} \mid X) = (X'\hat{\beta})^2 + \hat{V}(\varepsilon \mid X)$	$\sqrt{\hat{\varepsilon}} = X\theta + \eta_\varepsilon$ $\hat{V}(\varepsilon \mid X) = \left(X\cdot\hat{\theta}\right)^2 + MSE(\hat{\eta}_\varepsilon)$
GLM-scale	Log-gamma (scale) GLM	$E(\text{Cost} \mid X) = e^{(X'\beta)}$	$V(\text{Cost} \mid X) \propto \mu^2$
GLM-sqrt	Square-root-gamma (scale) GLM	$E(\text{Cost} \mid X) = (X'\beta)^2$	$V(\text{Cost} \mid X) \propto \mu^2$
GLM-shape	Log-gamma (shape) GLM	$E(\text{Cost} \mid X) = e^{(X'\beta)}$	$V(\text{Cost} \mid X) \propto \mu$
MP	Multipart Model: Individual parts use log-gamma (shape) GLMs; ordered logit used for mixing distribution	$E(\text{Cost} \mid X) = \sum_{K=1}^{4}\left(e^{(X'\beta_K)} \cdot P(K = k \mid X)\right)$	$V(\text{Cost} \mid X, K = k) \propto \mu$

Results

R^2, RMSE, and RMSE in the 10th Decile of predicted costs

Prediction Population	Estimation Cohorts	Result[b]	Model[a]								
			Original Nosos	SH	SL	SD	SS	GLM-sqrt	GLM-scale	GLM-shape	MP
GEC Overall	GEC Overall	R-square	0.54	0.61	0.61	0.62	0.61	0.59	0.48	0.61	0.63
		RMSE	43,904	40,428	41,094	39,884	40,201	41,123	46,563	40,034	39,449
		RMSE 10th Decile	106,102	99,945	100,960	97,958	99,164	101,921	122,797	97,807	95,957
	VA population	R-square	0.54	0.59	0.01	0.54	0.59	0.57	NEG[c]	0.56	0.61
		RMSE	43,904	41,688	65,260	43,790	41,282	42,490	201,108	43,259	40,630
		RMSE 10th Decile	106,102	102,423	187,179	111,188	102,233	103,616	627,626	105,040	99,962

[a] Original Nosos denotes the original Nosos model estimated only on the VA population; SH denotes the Square-root Transformed cost model with homoscedastic errors; SL denotes the Square-root Transformed cost model with log-transformed heteroskedastic errors; SD denotes the Square-root Transformed cost model with Duan heteroskedastic errors; SS denotes the Square-root Transformed cost model with square-root heteroskedastic errors; GLM-sqrt denotes the Square-root-gamma GLM; GLM-scale denotes the Log-gamma (scale) GLM; GLM-shape denotes the Log-gamma (shape) GLM; MP denotes the Multipart Model in which individual parts use log-gamma (shape) GLMs and an ordered logit is used for mixing distribution

[b] R-square denotes proportion of variance explained by the model; RMSE denotes the square root of the mean squared error; RMSE 10th Decile denotes the RMSE in the 10th prediction decile.

[c] NEG denotes a negative R^2, which indicates variation in model bias exceeds variation in the true conditional expectation

Consequences for evaluating program costs relative to standardized population

Prediction Population	Model[a]								
	Original Nosos	SH	SL	SD	SS	GLM-sqrt	GLM-scale	GLM-shape	MP
GEC, GeriPACT	−571	−3,208	−6,788	−2,117	−2,538	−1,637	−2,688	−4,311	−2,277
GEC, HCBS	2,390	571	−6,856	352	561	1,717	−1,963	572	737
GEC, HBPC	−4,928	−545	−8,154	−841	−597	48	−4,420	−346	−554

[a] Original Nosos denotes the original Nosos model estimated only on the VA population; SH denotes the Square-root Transformed cost model with homoscedastic errors; SL denotes the Square-root Transformed cost model with log-transformed heteroskedastic errors; SD denotes the Square-root Transformed cost model with Duan heteroskedastic errors; SS denotes the Square-root Transformed cost model with square-root heteroskedastic errors; GLM-sqrt denotes the Square-root-gamma GLM; GLM-scale denotes the Log-gamma (scale) GLM; GLM-shape denotes the Log-gamma (shape) GLM; MP denotes the Multipart Model in which individual parts use log-gamma (shape) GLMs and an ordered logit is used for mixing distribution

Conclusion

Multipart models can be a useful tool in the cost prediction toolbox.

Careful if want true representation of E($|X) rather than prediction

If we have time

Let's walk through my

program

Want the Stata program?

Contact me:

peter.veazie@va.gov

peter_veazie@urmc.rochester.edu

Acronyms

- GEC: Geriatrics & Extended Care
- VA: Department of Veterans Affairs
- ESRD: End-stage renal disease
- MLE: Maximum likelihood estimation
- QMLE: Quasi maximum likelihood estimation
- GLM: Generalized linear model

Thank you!