Choosing Models for Cost Analyses: Issues of Nonlinearity and Endogeneity

Melissa M. Garrido,[1,2] Partha Deb,[3]
James F. Burgess, Jr.,[4,5] Joan D. Penrod[1,2]

1. GRECC/REAP, James J Peters VA Medical Center, Bronx, NY
2. Brookdale Department of Geriatrics and Palliative Medicine, Mount Sinai School of Medicine
3. Department of Economics, Hunter College and the Graduate Center, City University of New York
4. Center for Organization, Leadership and Management Research, VA Boston Healthcare System
5. Department of Health Policy and Management, Boston University School of Public Health

Supported by Department Of Veteran Affairs, Health Services Research and Development Service (# IAD-06-060-2)

Cost Analyses and Comparative Effectiveness Research

"…The use and costs of health care are likely to be important outcomes of interventions for some patients, whether or not the results are used in a cost-effectiveness analysis."

Garber A, Sox H. *Health Affairs* 2010; 29(10): 1805-1811

Health spending accounted for 17.6% of the GDP in 2009

Question #1 (Poll): How comfortable do you feel performing health care cost analyses?

- Very comfortable
- Somewhat comfortable
- Neither comfortable nor uncomfortable
- Somewhat uncomfortable
- Very uncomfortable
- N/A – Do not perform health care cost analyses

Issues in Healthcare Cost Analyses

- Skewed data
- Non-negative outcomes
- Censoring
- Endogeneity

- Distribution of marginal effects

Question #2 (Whiteboard): Which methods do you use to analyze costs?

How to Analyze Costs with an Endogenous Regressor?

- Models that use instrumental variables
  - Two-stage least squares on cost (2SLS)
  - Two-stage least squares on the natural log of costs (log-2SLS)
  - Control function models (special case: two-stage residual inclusion [2SRI])
  - Full information maximum simulated likelihood (FIMSL)
- Propensity score models

Example using Veterans Health Administration Data:

Effect of an Inpatient Palliative Care Consultation on Costs

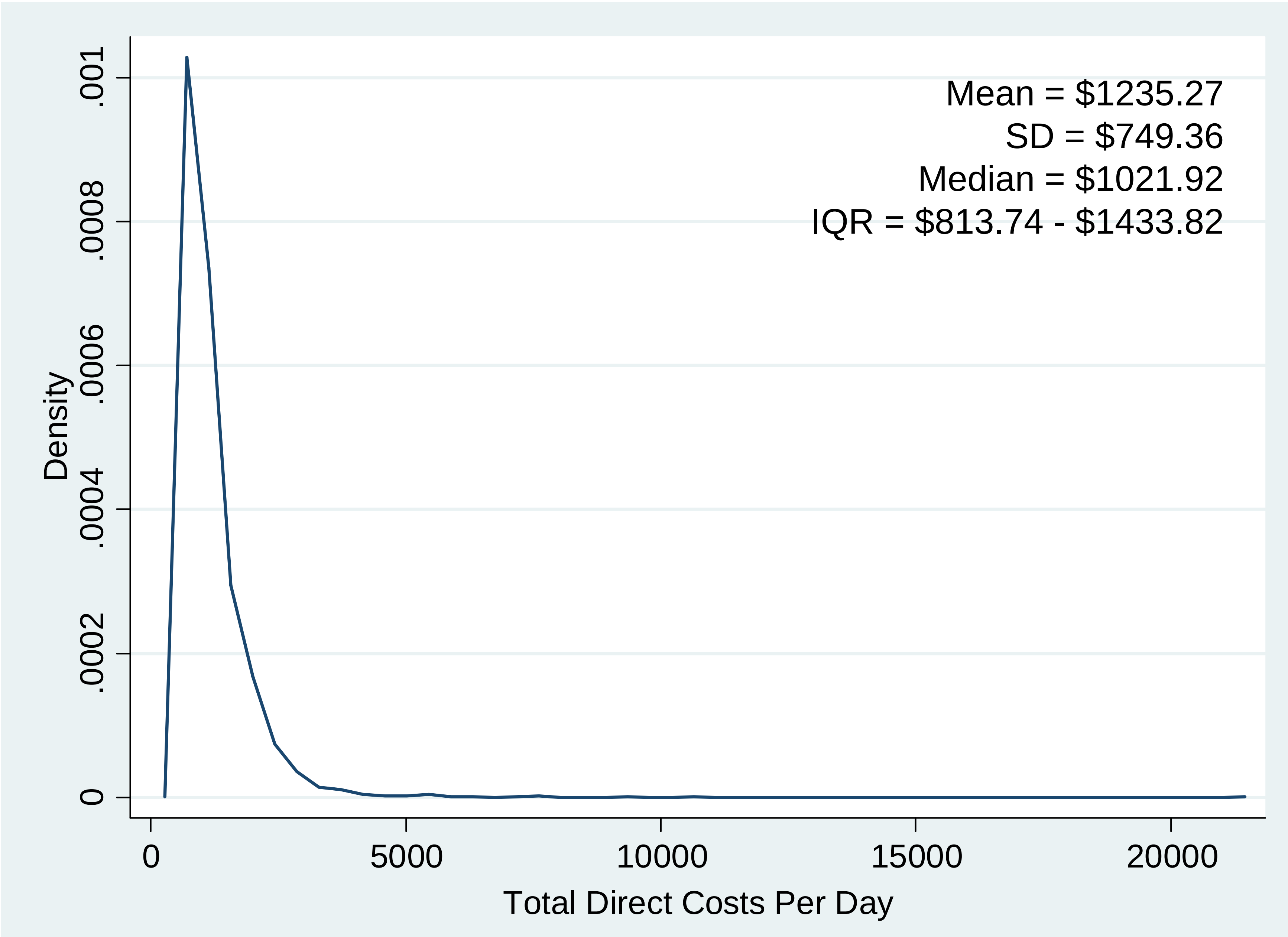
- 3,321 inpatients hospitalized in five Veterans Affairs acute care facilities in 2004-2007 with one or more life-limiting diseases
- Data from VHA Medical SAS Inpatient Dataset and VA Decision Support System National Data Extract
- Outcome: Total direct costs per day during the inpatient admission

Sample

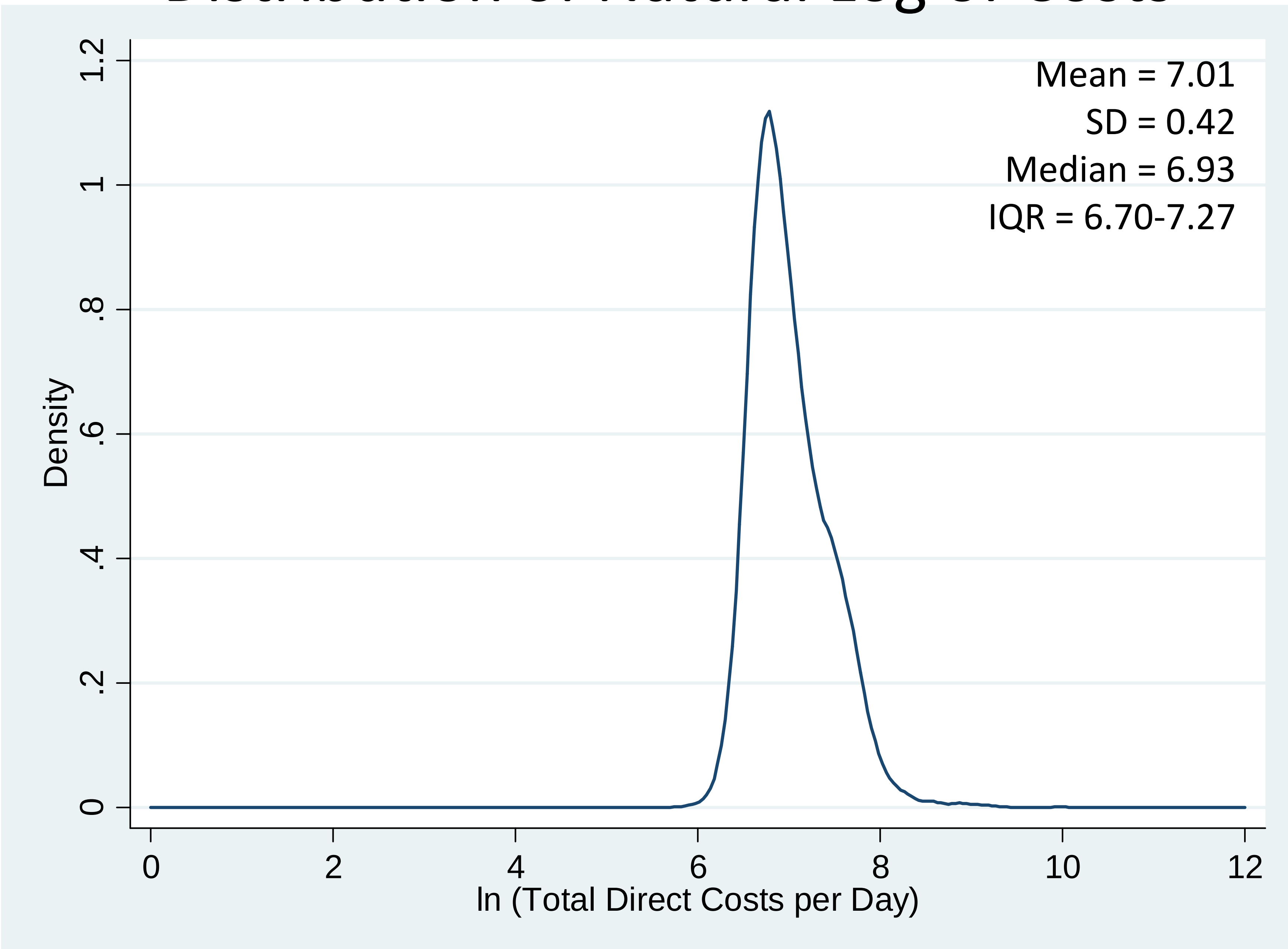
Variable	N (%) or Mean (SD)
Had a palliative care consultation	606 (18%)
65 or older	2235 (67%)
Advanced disease diagnosis*	
Cancer	802 (24%)
HIV/AIDS	54 (2%)
Congestive heart failure	1541 (46%)
Chronic obstructive pulmonary disease	1699 (51%)
Number of comorbidities at initial hospitalization	2.1 (1.3)
Died during study period	1659 (50%)
Cost per day	$1235.27 ($749.36), IQR $813.74-$1433.82
Natural log of costs per day	7.01 (0.42), IQR 6.70-7.27

* Could have more than one diagnosis

Skewed Distribution of Costs



Distribution of Natural Log of Costs



Instrumental variables

for equations with a continuous outcome and endogenous binary treatment

$$\Pr(\mathbf{d}_i = 1 \mid \mathbf{z}_i, \mathbf{I}_i) = \mathrm{g}(\mathbf{z}_i'\alpha + \boldsymbol{\delta}\mathbf{I}_i)$$

Distribution of error term

Treatment

Latent/Unobserved characteristics

Observed characteristics

$$E(Y_i \mid \mathbf{x}_i, \mathbf{d}_i, \mathbf{I}_i) = \mathrm{f}(\mathbf{x}_i'\beta + \gamma\mathbf{d}_i + \boldsymbol{\lambda}\mathbf{I}_i)$$

Distribution of error term

Observed characteristics

Outcome
(Cost)

Instrumental variables

for equations with a continuous outcome and endogenous binary treatment

$$\Pr(\mathbf{d}_i = 1 \mid \mathbf{z}_i, \mathbf{I}_i) = \mathrm{g}(\mathbf{z}_i^{'}\alpha + \boxed{\boldsymbol{\delta}\mathbf{I}_i})$$

$$E(Y_i \mid \mathbf{x}_i, \mathbf{d}_i, \mathbf{I}_i) = \mathrm{f}(\mathbf{x}_i^{'}\beta + \gamma\mathbf{d}_i + \boxed{\boldsymbol{\lambda}\mathbf{I}_i})$$

Treatment probability is endogenous due to common unobserved characteristics in each step (if neither δ or $\gamma = 0$)

Instrumental variables

for equations with a continuous outcome and endogenous binary treatment

$$\Pr(\mathbf{d}_i = 1 \mid \mathbf{z}_i, \mathbf{I}_i) = \mathrm{g}(\mathbf{z}_i^{'}\alpha + \boldsymbol{\delta}\mathbf{I}_i)$$

$$E(Y_i \mid \mathbf{x}_i, \mathbf{d}_i, \mathbf{I}_i) = \mathrm{f}(\mathbf{x}_i^{'}\beta + \gamma\mathbf{d}_i + \boldsymbol{\lambda}\mathbf{I}_i)$$

- 2SLS, Control Function, and Full Information Maximum Simulated Likelihood models are all based on these structural forms

- **z** must include at least one variable (the instrumental variable) that is not in **x**

- The instrumental variable must only be correlated with likelihood of treatment (**d**) and not with the likelihood of outcome (Y)

- Each uses a different set of assumptions about f and g

Instrumental Variable

- Treatment: Palliative care (PC) consultation
- Outcome: Healthcare costs
- Instrumental variable: Propensity for requesting a PC consultation by an admitting physician
  - First-stage F statistic = 19.46, $p < .001$
  - Chi-square value for the Anderson canonical correlation likelihood ratio test $p < .001$

2SLS

Step 1: Model treatment likelihood, include instrumental variable

Step 2: Model outcome likelihood, include treatment likelihood from Step 1

- Often used because:
  - Simple
  - Minimal assumptions on distribution of error term

- Problems:
  - Grossly inefficient
  - Misleading estimates when used with skewed outcomes (costs)

Log-2SLS

- 2SLS, but with natural log transformation of cost
- Problem: Retransformation of outcome back to costs to calculate marginal effects
  - Taking antilog of predicted value will lead to biased estimates
  - Smearing estimators work only in the case of homoskedastic error terms or when there are few, easily identifiable sources of heteroskedasticity

Manning WG. Journal of Health Economics. 1998. 17: 283-295

Control Functions

Step 1: Model treatment likelihood, include instrumental variable

Step 2: Model outcome likelihood, include a *function of the residuals* of the treatment likelihood equation

- Uses principles of instrumental variable regression
- Flexible function of the residuals can produce the correct adjustment for endogeneity in the outcome equation
- No need to transform dependent variable

Heckman JJ, Robb R. Journal of Econometrics 1985; 30: 329-267
Newey WK, Powell JL, Vella F. Econometrica 1999; 67(3): 565-603

Control Functions: Choice of Residuals

Types of Residuals

- Response
  - Difference between predicted and observed treatment likelihood
- Anscombe
  - Include transformations of the predicted and observed treatment likelihoods that are aimed at achieving normality
- Deviance
  - Obtained from a function of the log likelihood ratio

- These are equivalent in linear but not nonlinear settings
- Anscombe and deviance residuals better approximate a normal distribution in nonlinear settings

Gill, J. 2000. *Generalized Linear Models: A Unified Approach.* Thousand Oaks, CA: Sage.

Special Case of Control Function Approach: 2SRI

Step 1: Model treatment likelihood, include instrumental variable

Step 2: Model outcome likelihood, include the *response residual* of the treatment likelihood equation

Lee S. Journal of Econometrics 2007; 141: 1131-1158

Terza JV, Basu A, Rathouz PJ. Journal of Health Economics 2008; 27: 531-543

Often a misapplication of CF: No reason to believe that including just the response residual eliminates endogeneity bias

More research needed to determine which functional form and specification of the residuals is needed for CF to behave optimally in nonlinear studies

Basu A, Manning WG. Medical Care 2009; 47: S109-S114

Full Information Maximum Simulated Likelihood (FIMSL)

$$\Pr(\mathbf{d}_i = 1 \mid \mathbf{z}_i, \mathbf{I}_i) = \mathrm{g}(\mathbf{z}_i^{'}\alpha + \boldsymbol{\delta}\mathbf{I}_i)$$

$$E(Y_i \mid \mathbf{x}_i, \mathbf{d}_i, \mathbf{I}_i) = \mathrm{f}(\mathbf{x}_i^{'}\beta + \gamma\mathbf{d}_i + \boldsymbol{\lambda}\mathbf{I}_i)$$

- Derive the joint distribution of the treatment and outcome variables conditional on the common latent variables (I_i)
- Define and maximize a simulated likelihood function
- Estimate treatment effect with draws from a pseudo-random Halton sequence and average these effects

Deb P, Trivedi PK. Econometrics Journal. 2006; 9:307-331

Deb P, Trivedi PK. The Stata Journal. 2006; 6(2): 246-255

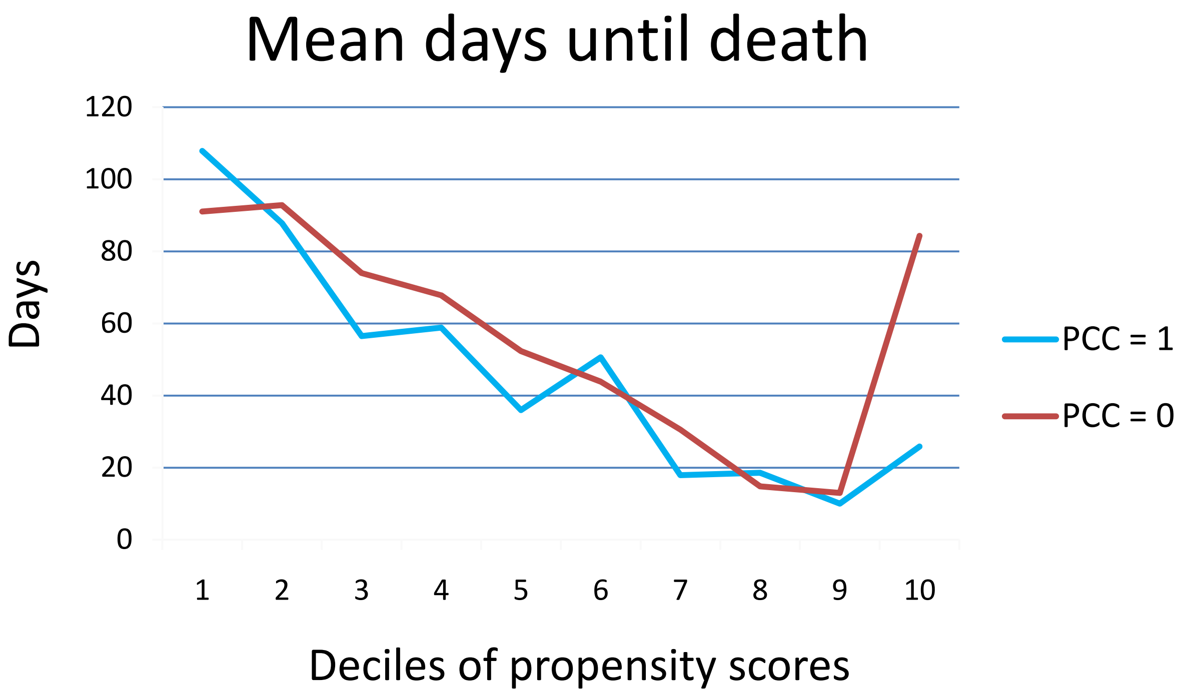
Propensity Scores

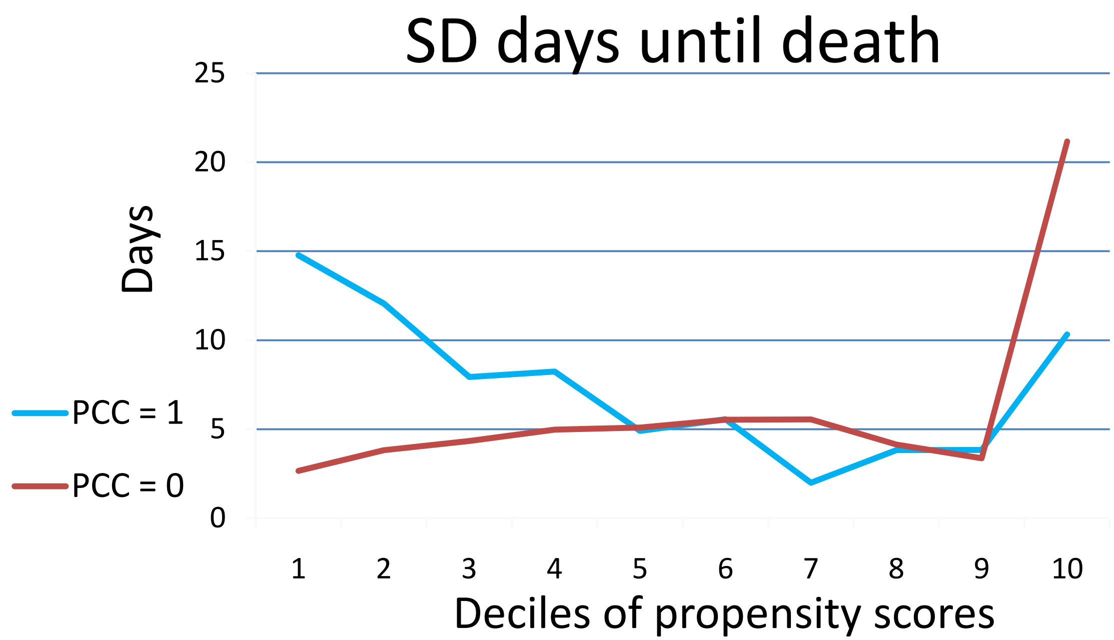
- Pros:
  - Accounts for selection into treatment based on observable characteristics

- Cons:
  - Little is known about the performance of propensity scores in nonlinear models
    - May not be able to balance distributions of covariates between those who did and did not receive treatment

Basu A, Manning WG. Medical Care 2009; 47 (7 Suppl 1): S109-S114

Basu A, Polsky D, Manning WG. Health, Econometrics and Data Group Working Paper 08/011. 2008


Mean days until death
120
100
80
60
40
20
0
Days
1
2
3
4
5
6
7
8
9
10
Deciles of propensity scores
PCC = 1
PCC = 0


SD days until death
25
20
15
10
5
0
Days
1
2
3
4
5
6
7
8
9
10
Deciles of propensity scores
PCC = 1
PCC = 0

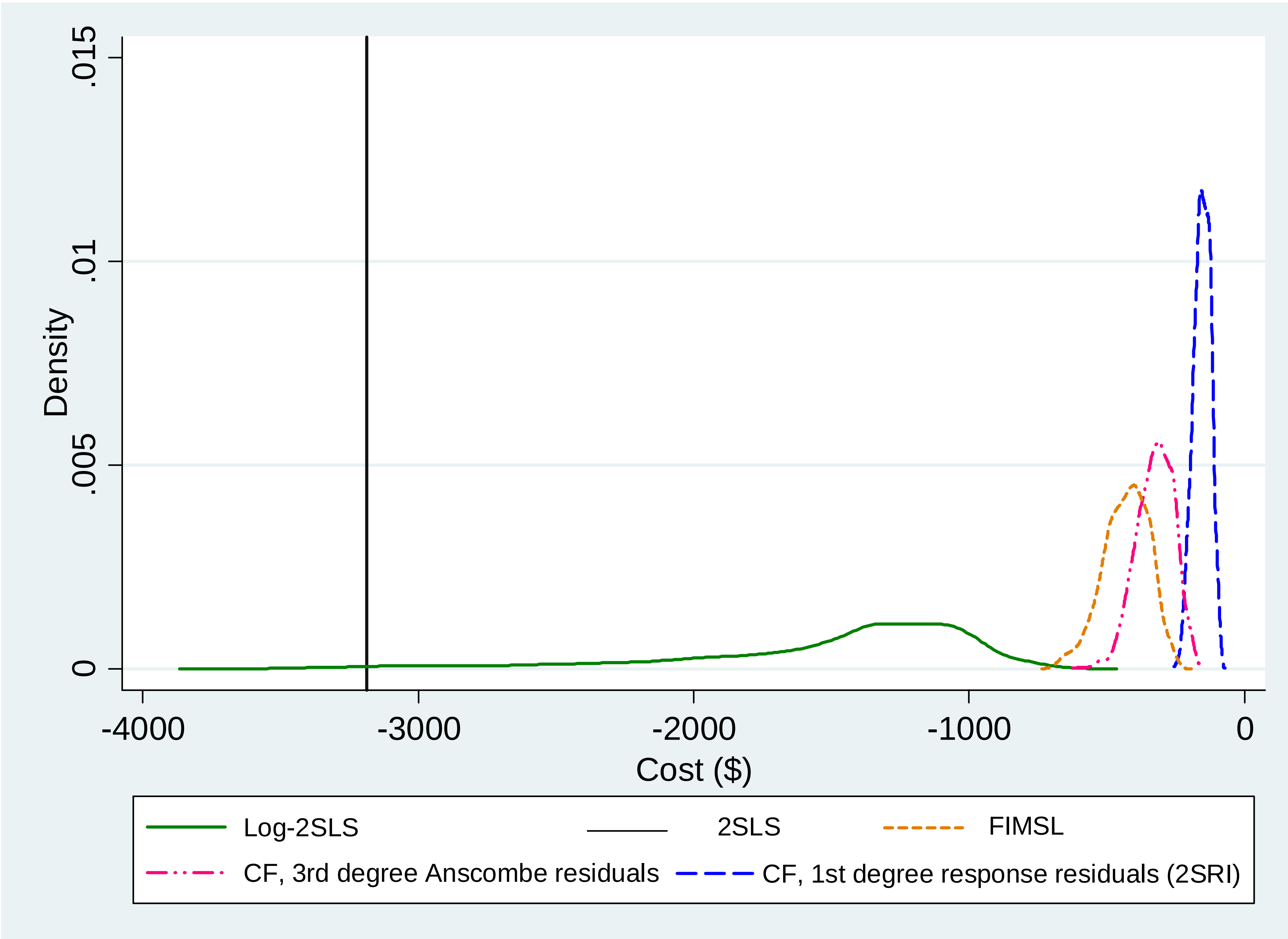
Model	Median	Mean	Range
Two Stage Least Squares			
Cost	-3185.83	-3185.83	(-3185.83, -3185.83)
Natural log of cost (retransformed)	-1347.77	-1514.07	(-3866.82, -464.14)
Propensity Score			
Nearest neighbor		-129.89	
Stratification		-95.97	
Control Function Approach			
Response – 1st degree (2SRI)	-152.33	-153.68	(-254.73, -70.95)
Response – 3rd degree	-217.64	-220.06	(-388.65, -100.49)
Anscombe – 1st degree	-75.29	-75.94	(-125.44, -35.76)
Anscombe – 3rd degree	-319.64	-324.62	(-625.05, -146.66)
Deviance – 1st degree	-133.34	-134.40	(-221.17, -62.47)
Deviance – 3rd degree	-439.39	-450.02	(-972.33, -198.25)
Full Information MSL	-422.92	-429.38	(-733.09, -193.93)

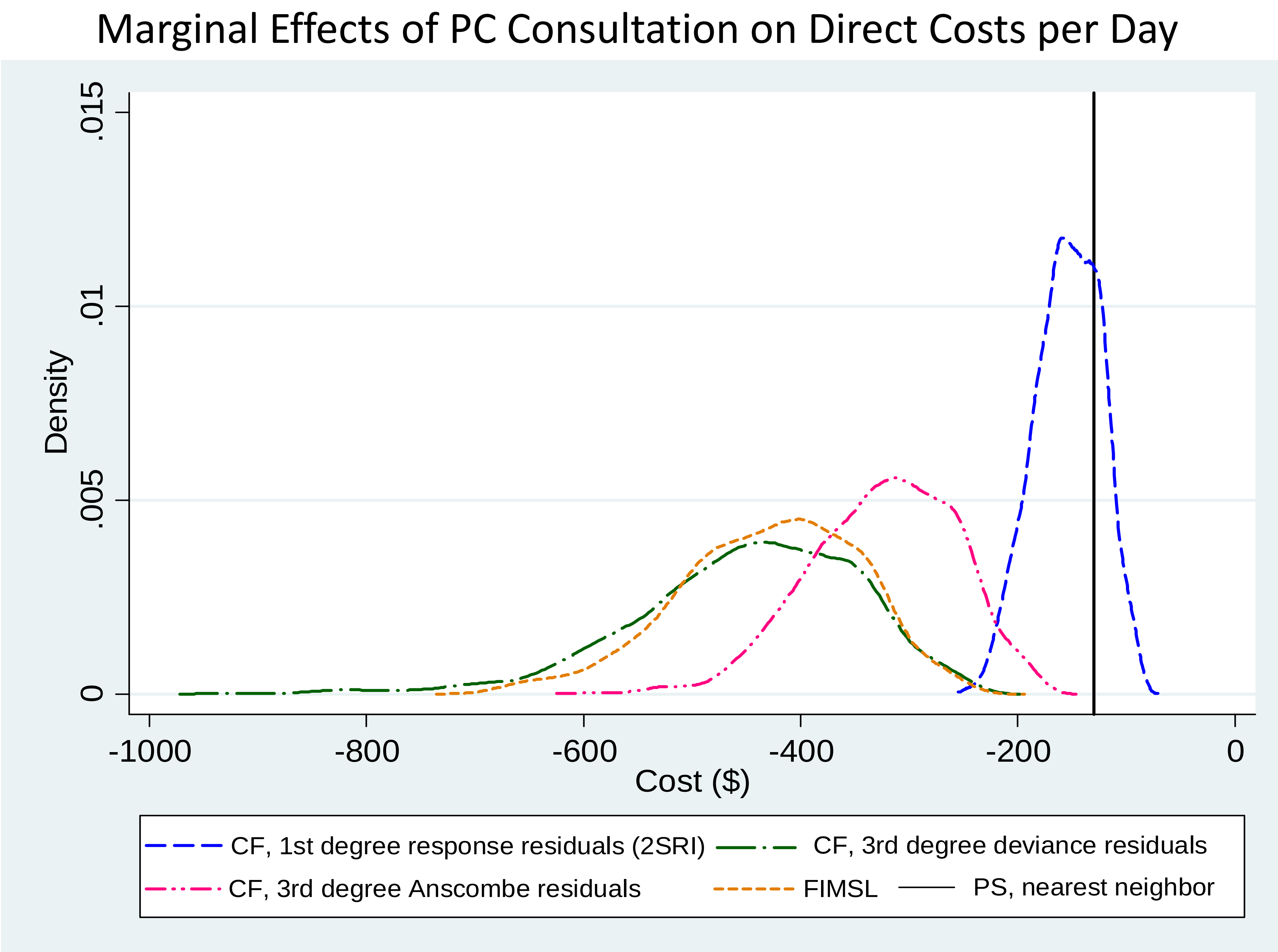
Importance of Distribution of Marginal Effects

- Partial effects differ across observations in nonlinear models

- Where and for whom the marginal effect is calculated depends on the research question

Marginal Effects of PC Consultation on Direct Costs per Day




Marginal Effects of PC Consultation on Direct Costs per Day
Density
0
.005
.01
.015
-1000
-800
-600
-400
-200
0
Cost ($)
CF, 1st degree response residuals (2SRI)
CF, 3rd degree deviance residuals
CF, 3rd degree Anscombe residuals
FIMSL
PS, nearest neighbor

Summary

- Estimates from the FIMSL model were the most similar to those of the CF 3rd degree Anscombe and deviance residuals

- Clustering of these three estimate distributions suggests that we have robust estimates of the effect of PC consultations on costs

- FIMSL provided similar results to CF without the need for several specification tests and with lower variance for estimates

- Further testing with other datasets is needed to determine how often FIMSL results mirror those from CF models

Recommendations for Health Care Cost Analyses

- Obtain estimates from several models to check for robustness of results
- Evaluate robustness by examining distributions of marginal effects
- Account for nonlinearity and endogeneity at the same time

Questions?

For list of references or other questions, please e-mail me at melissa.garrido@va.gov

The views expressed in this presentation are those of the authors and do not necessarily reflect the position or policy of the Department of Veterans Affairs or the United States government