

Analyzing Treatment Costs in Randomized Trials

Henry Glick
University of Pennsylvania

Society for Medical Decision Making Short Course #AM7

Phoenix, Az
October 20, 2012



Outline

- Part 1. Univariate analysis
 - Policy relevant parameter for CEA
 - Cost data 101
 - T-tests and response to the violation of normality
 - Primer on log cost
 - Why do different statistical tests lead to different inferences?
- Part 2. Multivariable analysis



Policy Relevant Parameter for CEA

- “Best estimate” of difference in population mean
 - In welfare economics, a project is cost-beneficial if winners from any policy gain enough to be able to compensate losers and still be better off themselves
 - Thus, we need a parameter that allows us to determine how much the losers lose, or cost, and how much the winners win, or benefit
 - From a budgetary perspective, decision makers can use arithmetic mean to determine how much they will spend on a program



Other Summary Statistics

- Summary statistics such as median cost may be useful in describing the data, but do not describe the difference in cost that will be incurred nor the cost saved by treating patients with one therapy versus another.
 - Not associated with social efficiency
- Lack of symmetry of cost distribution does not change the fact that we are interested in the arithmetic mean
- Even if evaluation of difference in medians or geometric satisfies the assumptions of the tests for these statistics, they generally do not answer the question we are asking

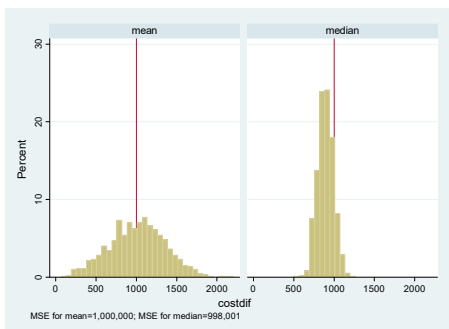


Truth in Advertising

- Does not mean median is never a better measure than the sample mean
- Relative bias – $(\text{observed difference} - \text{true difference})^2$ – generally used to determine better measure
- When cost data are sufficiently nonnormal, the relative bias for the median can be lower than the relative bias for the arithmetic mean



Relative Bias: $(\text{Observed} - \text{Truth})^2$



Are Sample Means Always the Best Estimator?

- Can be shown in simulation that when the log of cost is normally distributed, median has lower relative bias when the sample sizes are small and the true difference between the mean and median is small
- Given that in actual data we never know truth, it is difficult to determine whether other parameters will have lower relative bias than sample mean

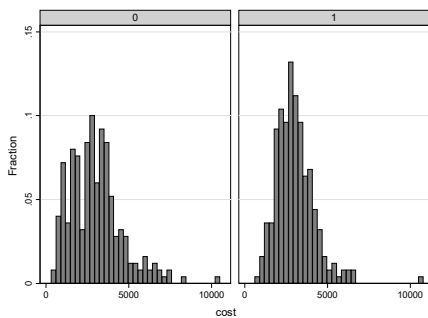


Cost Data 101

- Common feature of cost data is right-skewness (i.e., long, heavy, right tails)
- Data tend to be skewed because:
 - Can not have negative costs (but can have 0 cost)
 - Most severe cases may require substantially more services than less severe cases
 - Certain very expensive events occur in a relatively small number of patients
 - A minority of patients are responsible for a high proportion of health care costs



Typical Distribution Of Cost Data



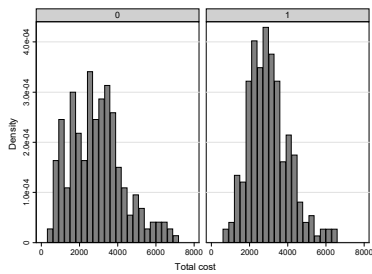
Typical Distribution Of Cost Data (II)

- Heavy tails vs. "outliers"
 - Distributions with long, heavy, right tails will have means that differ from the median
 - Median is not a better measure of the costs on average than is the mean



Problem Not Related Solely to "Outliers"

- Distribution when 5 observations with cost > 7000 (i.e., 3*SD) are eliminated



Mean, SD When 5 Observations with Cost > 7200 are Eliminated

	Full Sample		Trimmed (3*SD) *	
	Group 0	Group 1	Group 0	Group 1
Mean	3015	3040	2910	3010
Median	2826	2901	2813	2885

* p = 0.003 and 0.0001 for nonnormality of groups 0 and 1, respectively



Univariate Analysis: Parametric Tests Of Raw Means

- Usual starting point: T-tests and one way ANOVA
 - Used to test and estimate CI for differences in arithmetic means of total costs, QALYS, etc.
 - Makes assumption that the costs are normally distributed
 - Normality assumption is routinely violated for cost data, but t-tests shown to be robust to violations of this assumption when:
 - Samples moderately large
 - Samples are of similar size and skewness
 - Skewness is not too extreme



Responses To Violation Of Normality Assumption

- Adopt nonparametric tests of other characteristics of the distribution that are not as affected by the nonnormality of the distribution (“biostatistical” approach)
- Transform the data to approximate a normal distribution (“classic econometric” approach)
- Adopt tests of arithmetic means that avoid parametric assumptions



Response 1: Non-parametric Tests of Other Characteristics of the Distribution

- Rationale: Can analyze the characteristics that are not as affected by the nonnormality of the distribution
 - Wilcoxon rank-sum test
 - Test of difference in medians
 - Kolmogorov-Smirnov test
 - Test of difference in the cumulative distribution function



Potential Problem with Testing Other Characteristics of the Distribution

- Tests indicate that some measure of the cost distribution differs between the treatment groups, such as its shape or location, but not necessarily that the arithmetic means differ
- The resulting p-values need not be applicable to the arithmetic mean



Response 2: Transform the Data (I)

- Transform costs so they approximate a normal distribution
 - Common transformations
 - Log (arbitrary additional transformations required if any observation equals 0)
 - Square root
 - Estimate and draw inferences about differences in transformed costs




Estimates and Inferences Not Necessarily Applicable to Arithmetic Mean

- Goal is to use these estimates and inferences to estimate and draw inferences about differences in untransformed costs
 - Estimation: Simple exponentiation of the mean of log costs results in the geometric mean, which is a biased estimate of the arithmetic mean
 - Need to apply smearing factor
 - Inference: On the retransformed scale, inferences about the log of costs translate into inferences about differences in the geometric mean rather than the arithmetic mean




Log Transformation of Cost		
Raw Cost	Group 2	Group 3
Obs: 1	15	35
2	45	45
3	75	55
Arith mean	45	45
Log of arithmetic mean	3.806662	3.806662
Geometric mean $\sqrt[3]{\prod y}$	36.993	44.247
Log Cost		
Obs: 1	2.70805	3.55348
2	3.806662	3.806662
3	4.317488	4.007333
Arithmetic mean of logs	3.610734	3.789781
Exp ^(mean ln)	36.993	44.247



Primer On The Log Transformation Of Costs

- Exponentiation of mean of logs yields geometric mean
- In the presence of variability in costs, geometric mean is a downwardly biased estimate of the arithmetic mean
 - All else equal, the greater the variability, skewness, or kurtosis, the greater the downward bias
 - e.g., $(25 * 30 * 35)^{0.333} = 29.7196$
 - $(10 * 30 * 50)^{0.333} = 24.6621$
 - $(5 * 30 * 55)^{0.333} = 20.2062$
 - $(1 * 30 * 59)^{0.333} = 12.0664$
- "Smearing" factor attempts to eliminate bias from simple exponentiation of the mean of the logs




Retransformation Of The Log Of Cost (I)

- Duan's common smearing factor:

$$\Phi = \frac{1}{N} \sum_{i=1}^N e^{(Z_i - \hat{Z}_i)}$$

where in univariate analysis, \hat{Z}_i = the group mean

- Most appropriate when treatment group variances are equivalent



Retransformation Of The Log Of Cost (II)

Group	Observ	ln	$z_i - \bar{z}_i$	$e^{(z_i - \bar{z}_i)}$
2	1	2.708050	-0.9026834	0.4054801
2	2	3.806663	0.1959289	1.216440
2	3	4.317488	0.7067545	2.027401
Mean, 2	--	3.610734	--	--
3	1	3.555348	-0.2344332	0.7910191
3	2	3.806663	0.0168812	1.017025
3	3	4.007333	0.2175519	1.24303
Mean, 3	--	3.789781	--	--
Smear				1.116732



Common Smearing Retransformation (I)

- Retransformation formula

$$E(\hat{Y}_2) = \Phi e^{(z_2)}$$

$$E(\hat{Y}_3) = \Phi e^{(z_3)}$$

- Retransformation

Group	Φ	$e^{(\ln\hat{a}_i)}$	Predicted cost
2	1.116732	36.993	41.3
3	1.116732	44.247	49.4



Common Smearing Retransformation (II)

- Why are the retransformed subgroup-specific means -- 41.3 and 49.4 -- so different from the untransformed subgroup means of 45?
- Because the standard deviations of the subgroups' logs are substantially different
 - $SD_2 = 0.8224$; $SD_3 = 0.2265$
- The larger standard deviation for group 2 implies that compared with the arithmetic mean, its geometric mean has greater downward bias than does the geometric mean for group 3
- Thus, multiplication of the 2 groups' geometric means by a common smearing factor cannot yield accurate estimates for both groups' arithmetic means



Subgroup-specific Smearing Factors (I)

- Manning has shown that in the face of differences in variance – i.e., heteroscedasticity -- use of a common smearing factor in the retransformation of the predicted log of costs yields biased estimates of predicted costs
- We obtain unbiased estimates by use of subgroup-specific smearing factors
- Manning's subgroup-specific smearing factor:

$$\Phi_j = \frac{1}{N_j} \sum_{i=1}^{N_j} e^{(z_i - \hat{z}_j)}$$



Subgroup-specific Smearing Factors (II)

Group	Observ	ln	$z_i - \hat{z}_j$	$e^{(z_i - \hat{z}_j)}$
2	1	2.708050	-0.9026834	0.4054801
2	2	3.806663	0.1959289	1.216440
2	3	4.317488	0.7067545	2.027401
Mean, 2	--	3.610734	--	1.21644 Φ_2
3	1	3.555348	-0.2344332	0.7910191
3	2	3.806663	0.0168812	1.017025
3	3	4.007333	0.2175519	1.24303
Mean, 3	--	3.789781	--	1.0170245 Φ_3



Subgroup-specific Smearing Retransformation (I)

- Retransformation formulas

$$E(\bar{Y}_2) = \Phi_2 e^{\bar{z}_2}$$

$$E(\bar{Y}_3) = \Phi_3 e^{\bar{z}_3}$$

- Retransformation

Group	Φ_i	$e^{(\ln)}$	Predicted cost
2	1.21644	36.993	45.00
3	1.0170245	44.247	45.00



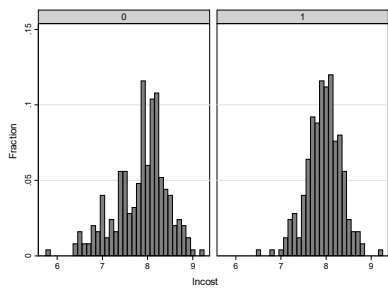
Subgroup-specific Smearing Retransformation (II)

- All else equal, in the face of differences in variance (or skewness or kurtosis), use of subgroup-specific smearing factors yield unbiased estimates of subgroup means
- Use of separate smearing factors eliminates efficiency gains from log transformation, because we cannot assume that p-value derived for the log of cost applies to the arithmetic mean of cost



Potential Problems with Testing Transformation of the Data (I)

- Log transformation doesn't always result in normality



P-value for normality = 0.002 and $p=0.01$ for the two groups



Potential Problems with Testing Transformation of the Data (II)

- P-value from t-test of log cost has direct applicability to the difference in the log of cost
- Generally also applies to the difference in the geometric mean of cost
 - Observe similar p-values for difference in log and difference in geometric mean
- P-value for log cost may or may not be directly applicable to difference in arithmetic mean of untransformed cost



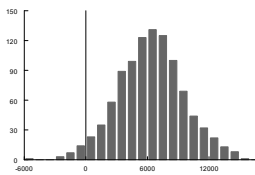
Potential Problems with Testing Transformation of the Data (III)

- Applicability of p-value for log to the difference in the arithmetic mean of untransformed cost depends on whether the two distributions of the log are normal and whether they have equal variance
 - If log cost is normally distributed and if the variances are equal, inferences about the difference in log cost are generally applicable to the difference in arithmetic mean cost
 - If log cost is normally distributed and if the variances are unequal, inferences about the difference in log cost generally will not be applicable to the difference in arithmetic mean cost



Response 3: Tests of Means that Avoid Parametric Assumptions

- Bootstrap estimates the distribution of the observed difference in arithmetic mean costs

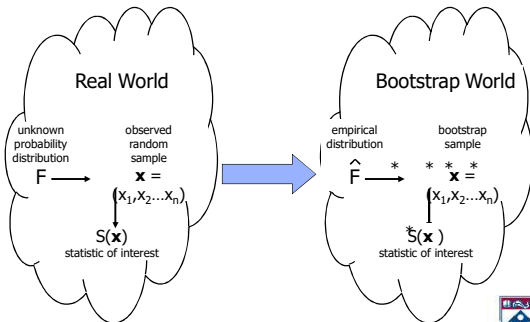


- Yields a test of how likely it is that 0 is included in this distribution (by evaluating the probability that the observed difference in means is significantly different from 0)



Bootstrap Simulation

Two worlds



Implementation of Bootstrap

- Random draw with replacement from each treatment group (thus creating multiple samples)
- Calculate the difference in the mean for each sample
- For the percentile method:
 - Count the number of replicates for which the difference is above and below 0 = 1-tailed p-value
 - Order and identify 2.5th and 97.5th percentile = 95% CI
- For parametric tests:
 - Because each bootstrap replicate represents a mean difference, when we sum the replicates, the reported "standard deviation" is the standard error
 - Difference in means / SE = t statistic
 - Difference in means + 1.96 SE = 95% CI



Example: Distribution of Costs, Chapter 5

	Group 0	Group 1
Arith Mean	3015	3040
Std. Dev.	1582.802	1168.737
Quantiles		
5%	899	1426
25%	1819	2226
50%	2825.5	2900.5
75%	3752	3604
95%	6103	5085
Skewness	1.03501	1.525386
Kurtosis	4.910192	9.234913
Geom Mean	2600.571	2835.971
Mean ln	7.8634864	7.9501397
SD ln	.57602998	.37871479
Obs	250	250

Data taken from Glick HA, Doshi JA, Sonnad SS, Polsky D. chapter 5 in Economic Evaluation in Clinical Trials, 2007.



Example: P Values from 6 Univariate Tests of the Difference in Cost

SUMMARY TABLE	P-value	95% CI
DIFFERENCE IN ARITHMETIC MEAN COST:	25.00	SE: 124.44
t-test, difference in means:	0.8409	-220 to 270
nonparametric BS, diff in means:	0.8600	-218 to 275
Wilcoxon rank-sum:	0.3722	
Kolmogorov-Smirnov:	0.0017	
t-test, difference in logs:	0.05	
transformation to normal:	Sqrt	
t-test, transformed variable:	0.2907	
test for heteroscedasticity:	0.0000	



Why Do Different Statistical Tests Lead To Different Inferences?

- The tests are evaluating differences in different statistics
 - T-test of untransformed costs indicates we cannot infer that the arithmetic means are different
 - Bootstrap leads to same (lack of) inference and does not make the normality assumption
 - Wilcoxon rank-sum test also leads to the same inference, but its p-value relates more to the probability that the medians differ
 - T-test of log costs indicates we can infer that the mean of the logs are different, and thus the geometric means of cost are different
 - Kolmogorov-Smirnov test indicates we can infer that the distributions are different



Outline for Part 2

- Part 1. Univariate analysis
- Part 2. Multivariable analysis
 - Ordinary least squares
 - Untransformed cost
 - Log of cost
 - General linear models (GLM)
 - Diagnostic tests



Multivariable Analysis Of Economic Outcomes (I)

- Even if treatment is assigned in a randomized setting use of multivariable analysis may have added benefits:
 - Improves the power for tests of differences between groups (by explaining variation due to other causes)
 - Facilitates subgroup analyses for cost-effectiveness (e.g., more/less severe; different countries/centers)
 - Variations in economic conditions and practice pattern differences by provider, center, or country may have a large influence on costs and the randomization may not account for all differences
 - Added advantage: Helps explain what is observed (e.g., coefficients for other variables should make sense economically)



Multivariable Analysis Of Economic Outcomes (II)

- If treatment is not randomly assigned, multivariable analysis is necessary to adjust for observable imbalances between treatment groups, but it may NOT be sufficient



Common Multivariable Techniques Used for the Analysis of Cost (I)

- Ordinary least squares regression predicting costs after randomization
- Ordinary least squares regression predicting the log transformation of costs after randomization
- Generalized Linear Models
- Other techniques:
 - Generalized Gamma regression (Manning et al., NBER technical working paper 293)
 - Extended estimating equations (Basu and Rathouz, Biostatistics 2005)



Problems with 'Typical' Methods

- Problems with OLS
 - Not robust
 - Can produce predictions with negative cost
- Problems with log OLS
 - Retransformation problem can lead to bias
 - Coefficients not directly interpretable
 - Residual may not be normally distributed even after log transformation
- More generally:
 - Assume $E(y/x) = \Sigma\beta_i X_i$ (OLS) OR $E(\ln(y)/x) = \Sigma\beta_i X_i$ (log OLS)
 - Assume constant variance



Generalized Linear Models (GLM)

- GLM models have the advantages of the log models, but
 - Don't require normality or homoscedasticity,
 - Evaluate the log of the mean, not the mean of the logs, and thus
 - Don't raise problems related to retransformation from the scale of estimation to the raw scale
- To build them, we must identify a "link function" and a "family" (based on the data)



GLM Relaxes OLS Assumptions

- Ability to choose among different links relaxes assumption that $E(y/x) = \sum \beta_i X_i$ (OLS) or $E(\ln(y)/x) = \sum \beta_i X_i$ (Log OLS)
- Ability to choose among different families relaxes assumption of constant variance
 - Gauss: constant variance
 - Poisson: variance proportional to mean
 - Gamma: variance proportional to square of mean
 - Inverse gauss: variance proportional to cube of mean



The Link Function

- Link function directly characterizes how the linear combination of the predictors is related to the prediction on the original scale
- While log link is most commonly used in literature, need not be the best fitting link
- Stata's power link provides a flexible link function
 - It allows generation of a wide variety of named and unnamed links, e.g.,
 - power 1 = Identity link = $B_i X_i$
 - power .5 = Square root link = $(B_i X_i)^2$
 - power 0 = log link = $\exp(B_i X_i)$
 - power -1 = reciprocal link = $1/(B_i X_i)$



The Log Link

- Log link is most commonly used in literature
- When we adopt the log link, we are assuming:

$$\ln(E(y/x)) = X\beta$$
- GLM with a log link differs from log OLS in part because in log OLS, we are assuming:

$$E(\ln(y)/x) = X\beta$$
- $\ln(E(y/x)) \neq E(\ln(y)/x)$
 i.e. log of the mean \neq mean of the log costs



$\ln(E(y/x)) \neq E(\ln(y)/x)$

Variable	Group 2	Group 3
Observations		
1	15	35
2	45	45
3	75	55
Arithmetic mean	45	45
Log, arith mean cost	3.806662	3.806662 *
Natural log		
1	2.70805	3.555348
2	3.806662	3.806662
3	4.317488	4.007333
Arithmetic mean	3.610734	3.789781 †

* Difference = 0; † Difference = 0.179047



Comparison of Results of GLM and log OLS Regression

Variable	Coefficient	SE	z/T	p value
GLM, gamma family, log link				
Group 3	0.000000	0.405730	0.00	1.000
Constant	3.806662	0.286894	13.27	0.000
Log OLS				
Group 3	0.179048	0.492494	0.36	0.74
Constant	3.610734	0.348246	10.32	0.000



Selecting a Link Function

- There is no single test that identifies the appropriate link
- Instead can employ multiple tests of fit
 - Pregibon link test checks linearity of response on scale of estimation
 - Modified Hosmer and Lemeshow test checks for systematic bias in fit on raw scale
 - Pearson's correlation test checks for systematic bias in fit on raw scale
- Ideally, all 3 tests will yield nonsignificant p-values



The Family

- Specifies the distribution that reflects the mean-variance relationship
 - Gaussian: Constant variance
 - Poisson: Variance is proportional to mean
 - Gamma: Variance is proportional to square of mean
 - Inverse Gaussian or Wald: Variance is proportional to cube of mean
- Use of the poisson, gamma, and inverse Gaussian families relax the assumption of homoscedasticity



Modified Park Test

- A "constructive" test that recommends a family given a particular link function
- Implemented after GLM regression that uses the particular link
- The test predicts the square of the residuals (res^2) as a function of the log of the predictions ($\ln\hat{y}$) by use of a GLM with a log link and gamma family to
 - Stata code
`glm res2 ln \hat{y} , link(log) family(gamma), robust`
- If weights or clustering are used in the original GLM, same weights and clustering should be used for modified Park test



Recommended Family, Modified Park Test

- Recommended family derived from the coefficient for $\ln y$ hat:
 - If coefficient $\sim=0$, Gaussian
 - If coefficient $\sim=1$, Poisson
 - If coefficient $\sim=2$, Gamma
 - If coefficient $\sim=3$, Inverse Gaussian or Wald
- Given the absence of families for negative coefficients:
 - If coefficient ≤ -0.5 , consider subtracting all observations from maximum-valued observation and rerunning analysis



Stata and SAS Code

- STATA code:
`glm y x, link(linkname) family (familyname)`
- General SAS code (not appropriate for gamma family / log link):

```
proc genmod;  
  model y=x/ link=linkname dist=familyname;  
run;
```



SAS Code for a Gamma Family / Log Link

- When running gamma/log models, the general SAS code drops observations with an outcome of 0
- If you want to maintain these observations and are predicting y as a function of x (M Buntin):

```
proc genmod;  
  a = _mean_;  
  b = _resp_;  
  d = b/a + log(a)  
  variance var = a2  
  deviance dev =d;  
  model y = x / link = log;  
run;
```



GLM Comments (I)

- Advantages
 - Relaxes normality and homoscedasticity assumptions
 - Consistent even if not the correct family distribution
 - Choice of family only affects efficiency if link function and covariates are specified correctly
 - Gains in precision from estimator that matches data generating mechanism
 - Avoids retransformation problems of log OLS models



GLM Comments (II)

- Disadvantages
 - Can suffer substantial precision losses
 - If heavy-tailed (log) error term, i.e., log-scale residuals have high kurtosis (>3)
 - If family is misspecified



Retransformation

- GLM avoids the problem that simple exponentiation of the results of log OLS yields biased estimates of predicted costs
- It does not avoid the other complexity of nonlinear retransformations (also seen in log OLS models):
 - On the transformed scale, the effect of the treatment group is estimated holding all else equal; however, retransformation (to estimate costs) reintroduces the covariate imbalances



Recycled Predictions

- Do not use the means of the covariates to avoid the reintroduction of covariate imbalance, because the mean of nonlinear retransformations does not equal the linear retransformation of the mean
- Rather, use the method of recycled predictions to create an identical covariate structure for the two groups by:
 - Coding everyone as if they were in treatment group 0 and predicting \hat{z}_0
 - Coding everyone as if they were in treatment group 1 and predicting \hat{z}_1



Extended Estimating Equations

- Basu and Rathouz (2005) have proposed use of extended estimating equations (EEE) which estimate the link function and family along with the coefficients and standard errors
- Tends to need a large number of observations (thousands not hundreds) to converge
- Currently can't take the results and use them with a simple GLM command (makes bootstrapping resulting models cumbersome)



Special Cases (I)

- A substantial proportion of observations have 0 costs
 - May pose problems to regression models
 - Commonly addressed by developing a “two-part” model in which the first part predicts the probability that the costs are zero or nonzero and the second part predicts the level of costs conditional on there being some costs
 - 1st part : Logit or probit model
 - 2nd part : log OLS or GLM model



Special Cases (II)

- Censored costs
 - Results derived from analyzing only the completed cases or observed costs are often biased
 - Need to evaluate the “mechanism” that led to the censored/missing data and adopt a method that gives unbiased results in the face of missingness



Which Statistic Should Be Used To Summarize Cost Data?

- Cost-effectiveness ratios ($\Delta C / \Delta E$) and NMB ($[WTP \Delta E] - \Delta C$) require an estimate of population differences in cost and effect
- Unless we are confident that some measure other than the sample mean is a better estimate of these differences in the population, we should evaluate differences in sample/arithmetical mean
 - Parametric test of means
 - Non-parametric test of means (e.g., bootstrap methods)



Multivariate Analysis: Summary/Conclusion

- Use mean difference in costs between treatment groups estimated from a multivariable model as the numerator for a cost-effectiveness ratio
- Establish criteria for adopting a particular multivariable model for analyzing the data prior to unblinding the data (i.e., the fact that one model gives a more favorable result should not be a reason for its adoption)
- Given that no method will be without problems, it may be helpful to report the sensitivity of our results to different specifications of the multivariable model