Analysis of Coronary Artery Calcification Data: Modeling Considerations

Nathaniel L Baker
Research Associate
DOM, Division of Biostatistics and Epidemiology
Tea Time for Science
4/11/2011
Analysis of Coronary Artery Calcification Data

Presentation Outline

1. Coronary Heart / Artery Disease
2. Coronary Artery Calcification
3. Analysis Methods
4. Simulation Study
5. Real Example and Results
6. Conclusions and Further work
The Disease
Coronary Heart / Artery Disease

The Disease

Coronary Artery Disease (CAD) is the leading cause of death in the US for both Men and women (NHLBI, 2009).

CAD is caused when the arteries that supply the heart with oxygenated blood become blocked by Plaque.

This condition is often referred to as Atherosclerosis and over time, can lead to heart attack, stroke, and death.
Atherosclerosis is caused when the inner lining of the arteries, specifically the endothelium, are injured or damaged.

Blood cells clump at the injury site in an attempt to repair the vessel wall, this leads to inflammation.

Plaques are then deposited on the artery wall and will continue to build over time.

These build ups rupture and harden over time and block the flow of blood.
Coronary Heart / Artery Disease

The Disease

Image Courtesy of NHLBI
The Disease

There are many factors that are linked to the development of Atherosclerosis (smoking, diet etc..).

Atherosclerosis is usually symptom free until severe blockages are present.

Early detection of coronary artery calcification (CAC) and narrowing is key to prevention of later events.

CAC has been shown to be an independent risk factor for cardiovascular events (Budoff et al, JACC 2007; Raggi et al, JACC 2004, Greenland et al, JAMA 2004).
Coronary Artery Calcification

Measurement of CAC

Ultra fast CT is used to detect and quantify CAC levels.

Measurements are usually given in AS (Agatston units), however it is sometimes measured as a volume (mm$^3$), or mass scores (Agatston, 1990).

Agatston Score measures the area of the plaque multiplied by some density factor. Scores can range from 0 to several thousand.
Coronary Artery Calcification

CAC Scores

N: 1205
Mean: 68.14878
Std Deviation: 254.0113
Minimum: 0
Maximum: 3101.325
Coronary Artery Calcification

CAC Scores

As you can see from the previous figure that raw CAC scores have a very high prevalence of 0 scores.

CAC scores are also notorious for being highly right skewed.

Historically, log(CAC+1) and Log(CAC) where CAC>0 have been used to facilitate linear regression techniques.

Other methods have also been implemented to assist in analysis and interpretation of CAC data.
Coronary Artery Calcification

The data above represents the distribution of log(CAC+1)…

while the data to the left is only the portion with measurable CAC
Current Analytic Methods
Analytic Challenges associated with CAC data.

Immeasurable CAC is present in many subjects and is represented by some lower bound (usually zero).

Large scores tend to violate many analytic assumptions.

Ordinary least squares regression analysis may be inappropriate and limited dependent variable data analysis can be complex and difficult to interpret.
Analysis Methods

Analytic Methods associated with CAC data.

1. Linear Regression
2. Restricted Linear Regression
3. Binary Logistic Regression
4. Multinomial Logistic Regression
5. Tobit Limited Dependent Regression
Linear regression model and assumptions

\[ y = \alpha + \beta x + \varepsilon \]

Linear Regression assumptions:

- Linear Relationship between x and y
- Independence of observations
- Normality of the error distribution \(\sim N(0, \sigma^2)\)
- Homoscedasticity
Why use linear regression

Estimates are easily interpreted and understood by clinicians.

Small sample sizes can be used.

Effect sizes are easily obtained.
Analysis Methods
Linear Regression

Why not to use linear regression

Censored data tends to produce inconsistent parameter estimates.

Estimation is sensitive to the normality of the error terms.

Deviations from normality and linearity can add substantial error to parameter estimates.

Non uniform variance due to censoring will cause standard error estimates to be either too small or too large.
Linear Regression assumptions

Linear Relationship between x and y:
there may be a linear relationship between observed CAC and the predictor, but not when the censored values are included

Constant error variance  Normality of the error
Analysis Methods
Linear Regression

Linear Regression Bias

Graciously borrowed from David Madigan’s web page
Analysis Methods
Restricted Linear Regression

Why don’t we just analyze the data that has measurable CAC values?

$$y = \alpha + \beta x + \varepsilon \text{ where } y \neq 0$$

We may satisfy the assumptions for linear regression analysis, but in most cases, 30-60% of subjects have non-measurable CAC.

Excluding these data points may add significant bias to the parameter estimates found in the model.
Analysis Methods
Logistic Regression

Logistic regression model and assumptions

\[
\text{logit}[\pi(x)] = \log \left( \frac{\pi(x)}{1 - \pi(x)} \right) = \alpha + \beta x
\]

\[
\pi(x) = \frac{e^{(\alpha + \beta x)}}{1 - e^{(\alpha + \beta x)}} \quad \text{Odds of CAC } > 0 = \frac{\pi(x)}{1 - \pi(x)} = e^{\alpha} (e^{\beta})^x
\]

Thus \( OR = e^\beta \)

Logistic Regression Assumptions:

- Underlying distribution is binary
- Independence of observations
- Linearity between IV’s and the Log odds
- Small sample sizes can produce poor power

Hosmer and Lemeshow (2000) recommend at least \( n=400 \)
Analysis Methods
Logistic Regression

**Why Use Logistic Regression?**

- The error terms do not have to be normally distributed
- The relationship between X and Y does not have to be linear
- There is no homogeneity of variance assumption
Why Not Use Logistic Regression?

Dichotomization can be harmful to estimation and hypothesis testing (Federov et al, 2009).

This leads to a loss of information/power and increased sample sizes to detect true effects.

If there is a non-linear effect, splitting the data will not allow detecting this relationship.

Requires much more data than OLS Regression, loss of information will cause an increase in sample size to maintain power.
Why Not Use Logistic Regression?

When the study is prospective; the incidence of measurable CAC is high and thus the odds ratio is an overestimate relative risk.

However, some have argued that since the distribution of CAC scores does not follow a known distribution, the information lost is minimal and the ease of results presentation makes up for the loss.
Analysis Methods
Multinomial / Ordinal Logistic Regression

Common classifications of CAC data

Reduce the continuous outcome CAC data to binary or ordinal response.

Some popular categorizations of CAC data…

Measureable (>0) vs. non measureable CAC (0)

Low CAC (≤10) vs. high CAC (>10)

Categories: 0, 0-10, 11-99, 100-399, 400-infinity

Have seen data driven cut points, but they are not recommended.
Analysis Methods
OLS and Logistic Regression

Problems with the approaches?

OLS regression is clearly inadequate in handling data with clustering at zero.

Binary regression models (logit, probit, LPM) are adequate if you are interested only in the probability of limit vs. non-limit responses. They fail to extract all of the information available.

Ordinal regression models with arbitrary cut points can be, but are rarely fully efficient.

Tobin (1958) proposed a latent model approach to deal with the zeros.
Truncated v Censored

Truncated: value is incomplete due to the selection process of the study. Usually occurs when both the dependent and independent variables are lost.

Censored: value is incomplete due to random factors for each subject. Usually occurs when data on the dependent variable is lost but not the independent variables. May be due to top / bottom coding.
Analysis Methods
Tobit Limited Dependent Regression

For left censored data, censored at $y_0$.

$$y^* = \alpha + \beta x + \varepsilon \quad \varepsilon \sim N(0, \sigma)$$

$$y = \begin{cases} 
    y^* & \text{if } y^* > y_0 \\
    0 & \text{if } y^* \leq y_0 
\end{cases}$$

The Tobit regression model assumes that the underlying dependent variable has negative values that are censored at zero. However, it is routinely used when observed values are clustered at zero, irrespective of censoring. (Sigelman and Zeng, 1999)
The log-likelihood function for the Tobit model when $y_0 = 0$:

$$\ln L = \sum_{i=1}^{N} \left\{ d_i \left( -\ln \sigma + \ln \phi \left( \frac{y_i - X_i \beta}{\sigma} \right) \right) + (1 - d_i)\ln \left( 1 - \Phi \left( \frac{X_i \beta}{\sigma} \right) \right) \right\}$$

There are two parts to the log-likelihood function.

Part 1:

$$d_i \left( -\ln \sigma + \ln \phi \left( \frac{y_i - X_i \beta}{\sigma} \right) \right)$$

This corresponds to the classical regression of uncensored variables.

Part 2:

$$(1 - d_i)\ln \left( 1 - \Phi \left( \frac{X_i \beta}{\sigma} \right) \right)$$

This corresponds to the relevant probabilities that an observation is censored.
The log-likelihood of the Tobit model is the sum of the log-likelihoods for each observation.

The Tobit model weights censored and uncensored values differently because of the log-likelihood function.

The Tobit model observes the censored values, but places more weight on the uncensored values for a more accurate estimate.

The OLS will weight every value equally, resulting in a poor model.
For CAC data, the Tobit model assumes that our data is censored at zero but may continue onto the negative scale if uncensored.

How can that be, an individual cannot have negative calcification? Can they?

No. But the distribution of CAC is a lognormal type and is transformed to conform to the assumptions of the model. When done, all of the CAC values less than 1 become negative log(CAC) values. So log(CAC) can take values that are negative. Note: log (CAC+1) values are always positive.
Why Use the Tobit Model

Handles the point mass zeros and the continuous data while producing a single parameter estimates.

Using OLS regression techniques will lead to downward biased and inconsistent parameter estimates.

Can be applied in most statistical software programs.
Why Not Use the Tobit Model

The Tobit censored regression model assumes that the error distribution of the underlying data is normal and is sensitive to violations of this assumption.

Heteroskedastic errors can lead to biased estimates where the OLS violation leads to underestimated standard errors.

Must graphically examine data to verify errors are i.i.d. normal.
Linear Regression: For a one unit change in the independent variable X, there is a $\hat{\beta}$ unit change in the dependent variable Y.

Restricted Linear Regression: For a one unit change in X, there is a $\hat{\beta}$ unit change in Y when Y>0.

Binary Logistic Regression: For a one unit change in X, there is a $\hat{\beta}$ unit change in the log odds of $Y_{\text{binary}}=1$ or for a one unit increase in X, the odds of $Y_{\text{binary}}=1$ increases by a factor of $e^{\hat{\beta}}$.

Ordinal Logistic Regression: For a one unit change in X, there is a $\hat{\beta}$ unit change in the log odds of $Y_{\text{ordinal}}$ being “higher”.
Tobit Censored Regression Model

Recall that in OLS there is only one conditional mean function

$$\frac{\partial E(y)}{\partial x_k} = \beta_k$$

The Tobit model has 3 conditional means (Greene, 1997)...

1. those of the latent variable $y^*$

$$\frac{\partial E(y^*|x)}{\partial x} = \beta$$

2. those of the observed dependent variable $y$

$$\frac{\partial E(y|x)}{\partial x} = \beta \Phi\left(\frac{x\beta}{\sigma}\right)$$

*Estimated probability of observing an uncensored event*
Tobit Censored Regression Model

The Tobit model has 3 conditional means (Greene, 1997)...

3. those of the uncensored observed dependent variable $y$

$$
\partial E(y \mid y > 0, x) / \partial x = \beta (1 - \delta \left( -\frac{x \beta}{\sigma} \right))
$$

Where

$$
\delta(\alpha) = \lambda(\alpha)(\lambda(\alpha) - 1)
$$

$$
\lambda(\alpha) = \phi(\alpha) / (1 - \Phi(\alpha))
$$

$$
\alpha = \left( \frac{x \beta}{\sigma} \right)
$$
Tobit Censored Regression Model

In most cases, software returns $\beta$ and is interpreted as the change in $x$ and its effect on $y^*$.  

**RECALL:** we are analyzing the logarithm of CAC, thus the parameter estimates of the linear regression models are the difference in logarithms $\approx$ logarithm of the ratio. We can exponentiate the estimate and CI and recover the ratio of geometric means which is roughly interpreted as the multiplicative increase in the true distribution of CAC for every unit change in the independent variable.
Simulation Study
The Goal of the simulation study is not to make statements based on the true distribution of CAC, rather to compare the performance of different analysis techniques with censored data.

In the study, the data is modeled such that the Tobit Censored regression model is correctly specified. The distribution of CAC conditioned on the covariates was normal with a uniform variance.

However, under normal conditions, CAC data may not have the properties desired.
Monte Carlo simulation done with 1000 samples sets of 1000 observations each...

\[
\log y^* = \alpha + \beta x + \varepsilon
\]

\[
\varepsilon \sim N(0,1) \quad x \sim N(100,4)
\]

The relationship between \(x\) and \(\log y^*\) is set to a known value of \(\beta = 1.0\).

Three different censoring patterns were examined, 25% left censored, 50% left censored, and 65% left censored.
### Simulation Study

The sample mean parameter estimates were noted and confidence intervals were calculated at the 2.5 and 97.5 percentiles

<table>
<thead>
<tr>
<th>Modeled Censoring</th>
<th>Continuous Models</th>
<th>Categorical Models</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>OLS</td>
<td>OLS (restricted)</td>
</tr>
<tr>
<td>25% left Censored</td>
<td>0.750 (0.717-0.783)</td>
<td>0.951 (0.927-0.975)</td>
</tr>
<tr>
<td>50% left Censored</td>
<td>0.499 (0.466-0.534)</td>
<td>0.906 (0.872-0.941)</td>
</tr>
<tr>
<td>65% left Censored</td>
<td>0.350 (0.317-0.383)</td>
<td>0.874 (0.827-0.922)</td>
</tr>
</tbody>
</table>

\[
\text{OLS}_\beta \approx \beta - (\text{OLS}_{\text{Bias}} \times \beta) \quad \text{When assumptions are met}
\]
What if the data assumptions are not met?

Austin et al (2000) presented a simulation study that compared OLS to Tobit regression in the presence of non normal error terms and non constant error variance.

When the Tobit model is correctly specified, the relative bias in the parameter estimate is close to 0 while the bias in the OLS model is proportional to % of censored observations.

When the conditional distribution is the mixture of 2 normal distributions, the bias in the tobit model remained < 10% even when the censoring % was high.
Simulation Study

What if the data assumptions are not met?

When the underlying data had a lognormal conditional distribution, again the OLS bias was approximately equal to the proportion of censoring while the Tobit parameter bias performed better with bias < 10%.

When the underlying data is normal with increasing variance, the Tobit model performs as poor or more poorly than the OLS model.

Lastly, when the underlying data is lognormal and the variance in increasing with X, the Tobit model again had a greater relative bias than the OLS model.
Example with real CAC data
The Diabetes Control and Complications Trial (DCCT) / Epidemiology of Diabetes Interventions and Complications (EDIC) study provides an opportunity to explore the complex relationships among traditional CVD risk factors, glycemia, and CVD outcomes.

As an example, we will examine the relationship between levels of CAC and the waist to hip ratio of each subject adjusted for other known covariates.
CT was performed on 1205 of the original 1441 subjects (84%) and 1189 have “natural waist to hip ratio” data available.

Data was analyzed by Cleary et al (2006), results were summarized using both logistic regression models (CAC=0, CAC>0) as well as the Tobit censored regression. Their focus was on metabolic memory and the effect of intensive treatment of diabetes on cardiovascular outcomes.

We will only focus on the a more simple age and gender adjusted analysis of the relationship between CAC and WH Ratio.
Results

Of the 1189 subjects with available CAC data, 821 have censored data (69%).

\[
\text{OLS}_\beta = \beta - (\text{OLS}_{\text{Bias}} \times \beta) \text{ thus } 1.19 - (1.19 \times 0.69) = 0.368 \approx 0.391. \text{ So, } 1.19 \text{ is not the true underlying Beta, but is probably close.}
\]
Conclusions and other work
Conclusions

Some basic conclusions

OLS Regression models will provide heavily biased estimated in the presence of censoring.

The Tobit regression model appears to be more robust in the presence of non normal data than OLS.

OLS performs better than the Tobit model in the presence of Heteroscedasticity.

Most published studies use a Logistic and/or Tobit regression modeling approach.
Conclusions

Tobit model in SAS

```sas
proc qlim data=cac;
  where (x ne .);
    model cac = x;
    endogenous cac ~ censored (lb=0);
    output out = fitted predicted expected conditional xbeta errstd;
run;

data fitted;
  set fitted;
  cdf = probnorm (xbeta_cac/ errstd_cac);
  pdf = PDF('NORMAL',xbeta_cac/errstd_cac);
  y_censored_expected = cdf * xbeta_cac + errstd_cac * pdf;
  /* This is E(Y|X)*/
run;

The Tobit model can also be implemented in SAS Proc Lifereg, R (VGAM), STATA (tobit), and Mplus
```
Further Work

Models used in the literature

Tobit Regression, Logistic regression, OLS Regression, Probit Regression, Risk Regression, Median Test, Generalized Additive Models

Other Suggestions

Two Part Models
   Logit-linear & Probit-linear, Han and Kronmal, 2006
   Probit/Log Skew normal, Chai and Bailey, 2008

*implemented using MLE in proc nlmixed*
References

Collaborators

Maria F Lopes-Virella, MD PhD
MUSC, Endocrinology, Diabetes and Medical Genetics

Kelly J Hunt, PhD
MUSC, Biostatistics and Epidemiology

John M Lachin, ScD
GWU, Biostatistics, Epidemiology and Statistics

David Nathan, MD
Harvard Medical School, Medicine

Gabe Virella, MD PhD
MUSC, Microbiology and Immunology

The DCCT/EDIC Research Group
Further Work

Is the underlying distribution of log(CAC) truly symmetric and normal?

If not, how biased will the estimates become and how will the methods compare?

Fleishman’s Power Transformation Method can be used to add varying levels of skew and kurtosis to the distribution and retest the models noting the bias. (Fleishman, 1978)