Copula Regression

RAHUL A. PARSA
DRAKE UNIVERSITY

&

STUART A. KLUGMAN
SOCIETY OF ACTUARIES

CASUALTY ACTUARIAL SOCIETY
MAY 18,2011

Outline

- Ordinary Least Squares (OLS) Regression
- Generalized Linear Models (GLM)
- Copula Regression
  - Continuous case
  - Discrete Case
- Examples

Notation

- Notation:
- Y – Dependent Variable
- $X_1, X_2, \cdots X_k$ Independent Variables
- Assumption
- Expected value of Y is related to X's in some functional form

$$\mathrm{E}[Y \mid X_1 = x_1, \ldots, X_n = x_n] = f(x_1, x_2, \ldots, x_n)$$

OLS Regression

- The Ordinary Least Squares model has Y linearly dependent on the Xs.

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \varepsilon_i$$
$$\varepsilon_i \sim \text{Normal}(0, \sigma^2) \text{ and independent}$$

OLS Regression

- The parameter estimate can be obtained by least squares. The estimate is:

$$\hat{Y} = (X'X)^{-1} X'y$$
$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \cdots + \hat{\beta}_k x_{ki}$$

OLS - Multivariate Normal Distribution

- Assume $Y, X_1, \ldots, X_k$ jointly follow a multivariate normal distribution. This is more restrictive than usual OLS.
- Then the conditional distribution of Y | **X** has a normal distribution with mean and variance given by

$$E(Y \mid \underset{\sim}{X} = \underset{\sim}{x}) = \underset{\sim}{\mu}_y + \Sigma_{YX} \Sigma_{XX}^{-1} (\underset{\sim}{x} - \underset{\sim}{\mu}_x)$$
$$\text{Variance} = \Sigma_{YY} - \Sigma_{YX} \Sigma_{XX}^{-1} \Sigma_{YX}$$

OLS & MVN

- Y-hat = Estimated Conditional mean
- It is the MLE
- Estimated Conditional Variance is the error variance
- OLS and MLE result in same values
- Closed form solution exists

Generalization of OLS

- Is *Y* always linearly related to the *X*s?
- What do you do if the relationship between is non-linear?

GLM – Generalized Linear Model

- $Y|x$ belongs to the exponential family of distributions and
  $E(Y \mid \underset{\sim}{X} = \underset{\sim}{x}) = g^{-1}(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)$
- g is called the link function
- *x*s are not random
- Conditional variance is no longer constant
- Parameters are estimated by MLE using numerical methods

GLM

- Generalization of GLM: Y can have any conditional distribution (See *Loss Models*)
- Computing predicted values is difficult
- No convenient expression for the conditional variance

Copula Regression

- Y can have any distribution
- Each X_i can have any distribution
- The joint distribution is described by a Copula
- Estimate Y by $E(Y|\boldsymbol{X}=\boldsymbol{x})$ – conditional mean

Copula

Ideal Copulas have the following properties:

- ease of simulation
- closed form for conditional density
- different degrees of association available for different pairs of variables.

Good Candidates are:

- **Gaussian or MVN Copula**
- t-Copula

MVN Copula -cdf

- CDF for the MVN Copula is

$$F(x_1, x_2, \ldots, x_n) = G(\Phi^{-1}[F(x_1)], \ldots, \Phi^{-1}[F(x_n)])$$

- where G is the multivariate normal cdf with zero mean, unit variance, and correlation matrix R.

MVN Copula - pdf

- The density function is

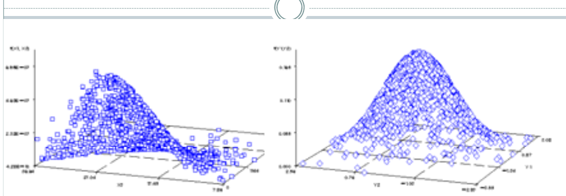
$$f(x_1, x_2, \ldots, x_n)$$

$$= f(x_1) f(x_2) \cdots f(x_n) \exp\left\{-\frac{v^T (R^{-1} - I) v}{2}\right\} * |R|^{-0.5}$$

Where v is a vector with ith element

$$v_i = \Phi^{-1}[F(x_i)]$$

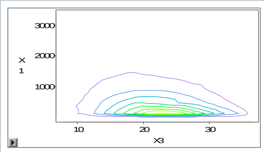
Copula vs. Normal Density



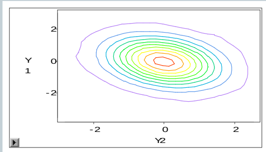
Bivariate Normal Copula with Beta and Gamma marginals

Bivariate Normal Distribution

Copula vs. Normal



Contour plot of the Bivariate Normal Copula with Beta and Gamma marginals



Contour plot of the Bivariate Normal Distribution

Conditional Distribution in MVN Copula

- The conditional distribution is

$$f(x_n \mid x_1,\ldots,x_{n-1})$$

$$= f(x_n)\exp\left\{-0.5\left[\frac{\{\Phi^{-1}[F(x_n)]-r^T R_{n-1}^{-1} v_{n-1}\}^2}{(1-r^T R_{n-1}^{-1} r)} - \{\Phi^{-1}[F(x_n)]\}^2\right]\right\}$$

$$\times(1-r^T R_{n-1}^{-1} r)^{-0.5}$$

$$v_{n-1} = (v_1,\ldots,v_{n-1}) \qquad R = \begin{bmatrix} R_{n-1} & r \\ r^T & 1 \end{bmatrix}$$

Copula Regression - Continuous Case

- Parameters are estimated by MLE.
- If $Y, X_1,\ldots, X_k$ are continuous variables, then we can use the previous equation to find the conditional mean.
- One-dimensional numerical integration is needed to compute the mean.

Copula Regression -Discrete Case

When one of the covariates is discrete

Problem:

- Determining discrete probabilities from the Gaussian copula requires computing many multivariate normal distribution function values and thus computing the likelihood function is difficult.

Copula Regression – Discrete Case

Solution:

- Replace discrete distribution by a continuous distribution using a uniform kernel.

Copula Regression – Standard Errors

- How to compute standard errors of the estimates?
- As n -> ∞, the MLE converges to a normal distribution with mean equal to the parameters and covariance the inverse of the information matrix.

$$I(\theta) = -n * E\left[\frac{\partial^2}{\partial \theta^2} \ln(f(X,\theta))\right]$$

How to compute Standard Errors

- *Loss Models*: "To obtain the information matrix, it is necessary to take both derivatives and expected values, which is not always easy. A way to avoid this problem is to simply not take the expected value."
- It is called "Observed Information."

Examples

- All examples have three variables – simulated using MVN copula
- R Matrix :

1	0.7	0.7
0.7	1	0.7
0.7	0.7	1

- Error measured by $\sum(Y_i - \hat{Y}_i)^2$
- Also compared to OLS

Example 1

- Dependent – Gamma; Independent – both Pareto
- X2 did not converge, used gamma model

Variables	X1-Pareto	X2-Pareto	X3-Gamma
Parameters	3, 100	4, 300	3, 100
MLE	3.44, 161.11	1.04, 112.003	3.77, 85.93

Error:

Copula	59000.5
OLS	637172.8

Example 1 - Standard Errors

- Diagonal terms are standard deviations and off-diagonal terms are correlations

	X_1 Pareto		X_2 Gamma		X_3 Gamma				
	$Alpha_1$	$Theta_1$	$Alpha_2$	$Theta_2$	$Alpha_3$	$Theta_3$	R(2,1)	R(3,1)	R(3,2)
$Alpha_1$	0.266606	0.966067	0.359065	-0.33725	0.349482	-0.33268	-0.42141	-0.33863	-0.29216
$Theta_1$	0.966067	15.50974	0.390428	-0.25236	0.346448	-0.26734	-0.37496	-0.29323	-0.25393
$Alpha_2$	0.359065	0.390428	0.025217	-0.78766	0.438662	-0.35533	-0.45221	-0.30294	-0.42493
$Theta_2$	-0.33725	-0.25236	-0.78766	3.558369	-0.38489	0.464513	0.496853	0.35608	0.470009
$Alpha_3$	0.349482	0.346448	0.438662	-0.38489	0.100156	-0.93602	-0.34454	-0.46358	-0.46292
$Theta_3$	-0.33268	-0.26734	-0.35533	0.464513	-0.93602	2.485305	0.365629	0.482187	0.481122
R(2,1)	-0.42141	-0.37496	-0.45221	0.496853	-0.34454	0.365629	0.010085	0.457452	0.465885
R(3,1)	-0.33863	-0.29323	-0.30294	0.35608	-0.46358	0.482187	0.457452	0.01008	0.481447
R(3,2)	-0.29216	-0.25393	-0.42493	0.470009	-0.46292	0.481122	0.465885	0.481447	0.009706

Example 1

- Maximum likelihood estimate of correlation matrix

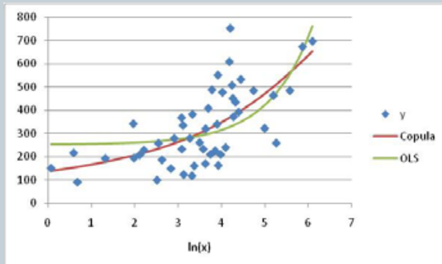
R-hat =

1	0.711	0.699
0.711	1	0.713
0.699	0.713	1

Example 1a – Two dimensional

- Only X3 (dependent) and X1 used.
- Graph on next slide (with log scale for x) shows the two regression lines.

Example 1a - Plot



Example 2

- Dependent – X3 - Gamma
- X1 & X2 estimated empirically (so no model assumption made)

Variables	X1-Pareto	X2-Pareto	X3-Gamma
Parameters	3, 100	4, 300	3, 100
MLE	F(x) = x/n – 1/2n f(x) = 1/n	F(x) = x/n – 1/2n f(x) = 1/n	4.03, 81.04

Error:

Copula	595,947.5
OLS	637,172.8
GLM	814,264.754

Example 2 – empirical model

- As noted earlier, when a marginal distribution is discrete MVN copula calculations are difficult.
- Replace each discrete point with a uniform distribution with small width.
- As the width goes to zero, the results on the previous slide are obtained.

Example 3

- Dependent – X3 – Gamma
- X1 has a discrete, parametric, distribution
- Pareto for X2 estimated by Exponential

Variables	X1-Poisson	X2-Pareto	X3-Gamma
Parameters	5	4, 300	3, 100
MLE	5.65	119.39	3.67, 88.98

- Error:

Copula	574,968
OLS	582,459.5

Example 4

- Dependent – X3 - Gamma
- X1 & X2 estimated empirically
- C = # of obs ≤ x and a = (# of obs = x)

Variables	X1-Poisson	X2-Pareto	X3-Gamma
Parameters	5	4, 300	3, 100
MLE	F(x) = c/n + a/2n f(x) = a/n	F(x) = x/n – 1/2n f(x) = 1/n	3.96, 82.48

Error:

Copula	OLS	GLM
559,888.8	582,459.5	652,708.98

Example 4 – discrete marginal

- Once again, a discrete distribution must be replaced with a continuous model.
- The same technique as before can be used, noting that now it is likely that some values appear more than once.

Example 5

- Dependent – X1 - Poisson
- X2, estimated by exponential

Variables	X1-Poisson	X2-Pareto	X3-Gamma
Parameters	5	4, 300	3, 100
MLE	5.65	119.39	3.66, 88.98

Error:

Copula	108.97
OLS	114.66

Example 6

- Dependent – X1 - Poisson
- X2 & X3 estimated empirically

Variables	X1-Poisson	X2-Pareto	X3-Gamma
Parameters	5	4, 300	3, 100
MLE	5.67	F(x) = x/n – 1/2n f(x) = 1/n	F(x) = x/n – 1/2n f(x) = 1/n

Error:

Copula	110.04
OLS	114.66