Five Things You Should Know about Quantile Regression

Phil Gibbs
SAS Technical Support

Quantile regression brings the familiar concept of a percentile into the framework of linear models

Goal

Interpretability and accurate prediction

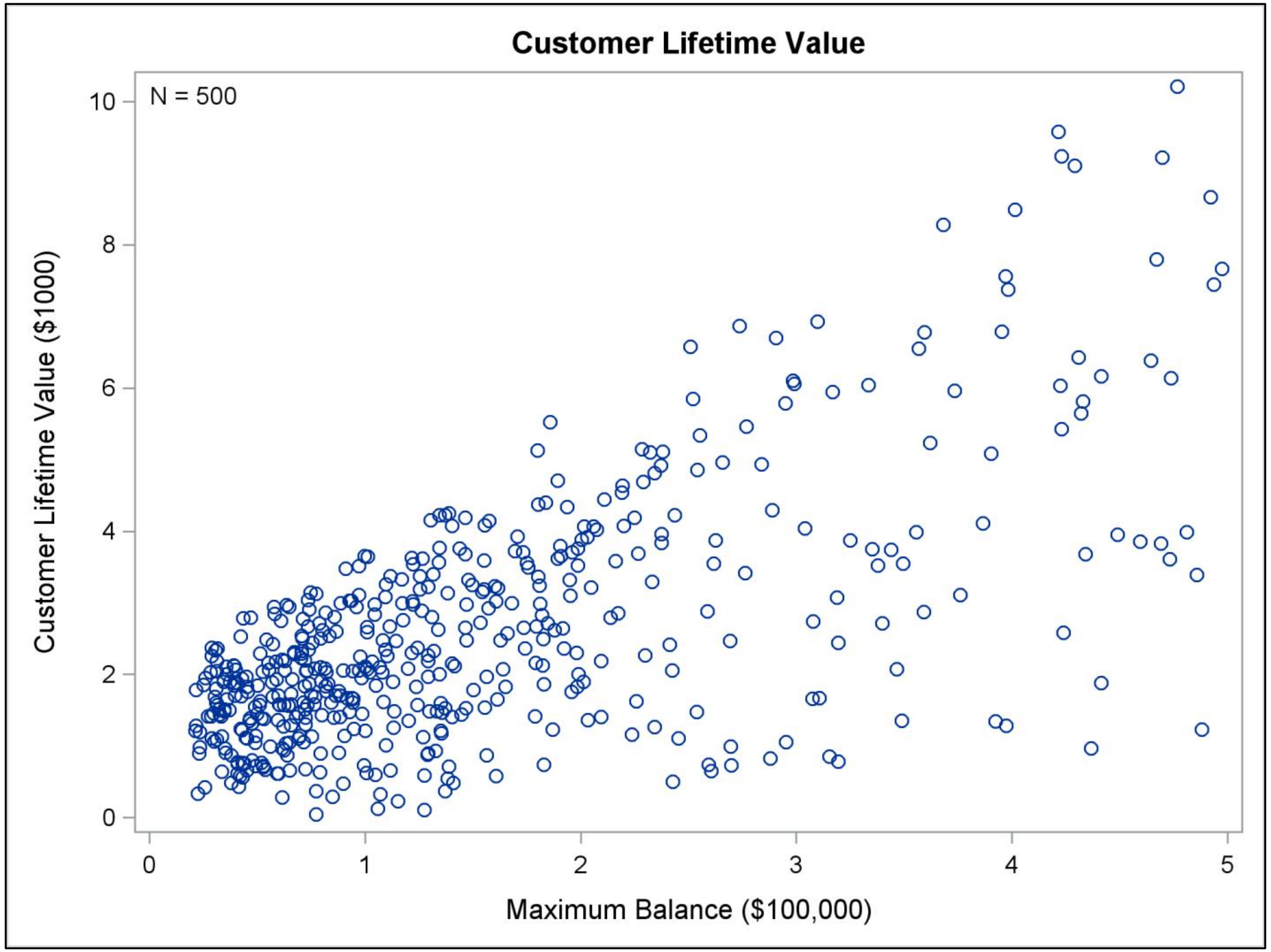
$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i \ , \quad i = 1, \ldots, n$$

Outline

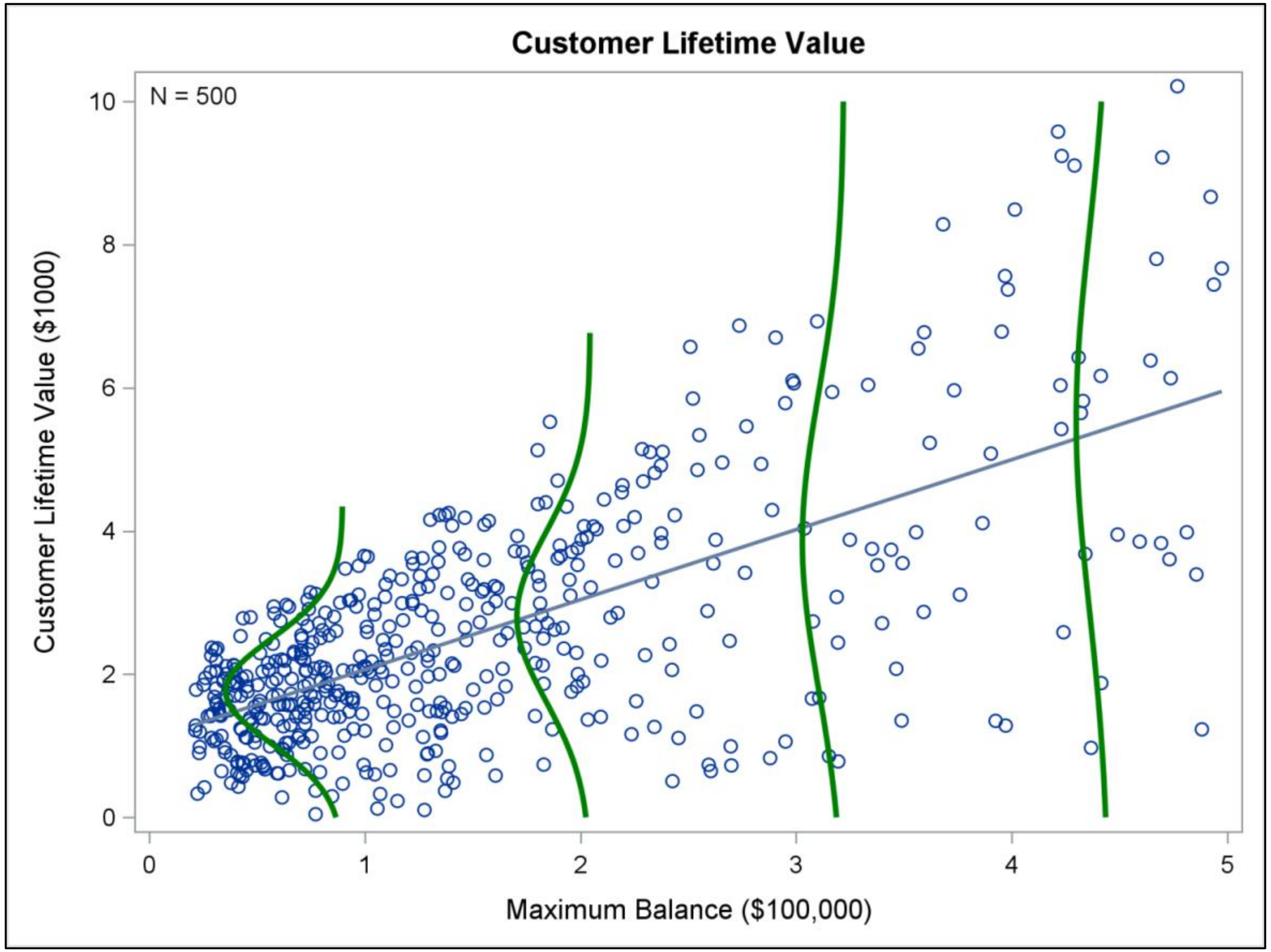
- Basic concepts
- Fitting and building quantile regression models
- Application to risk management
- Application to ranking student exam performance

Basic Concepts of Quantile Regression

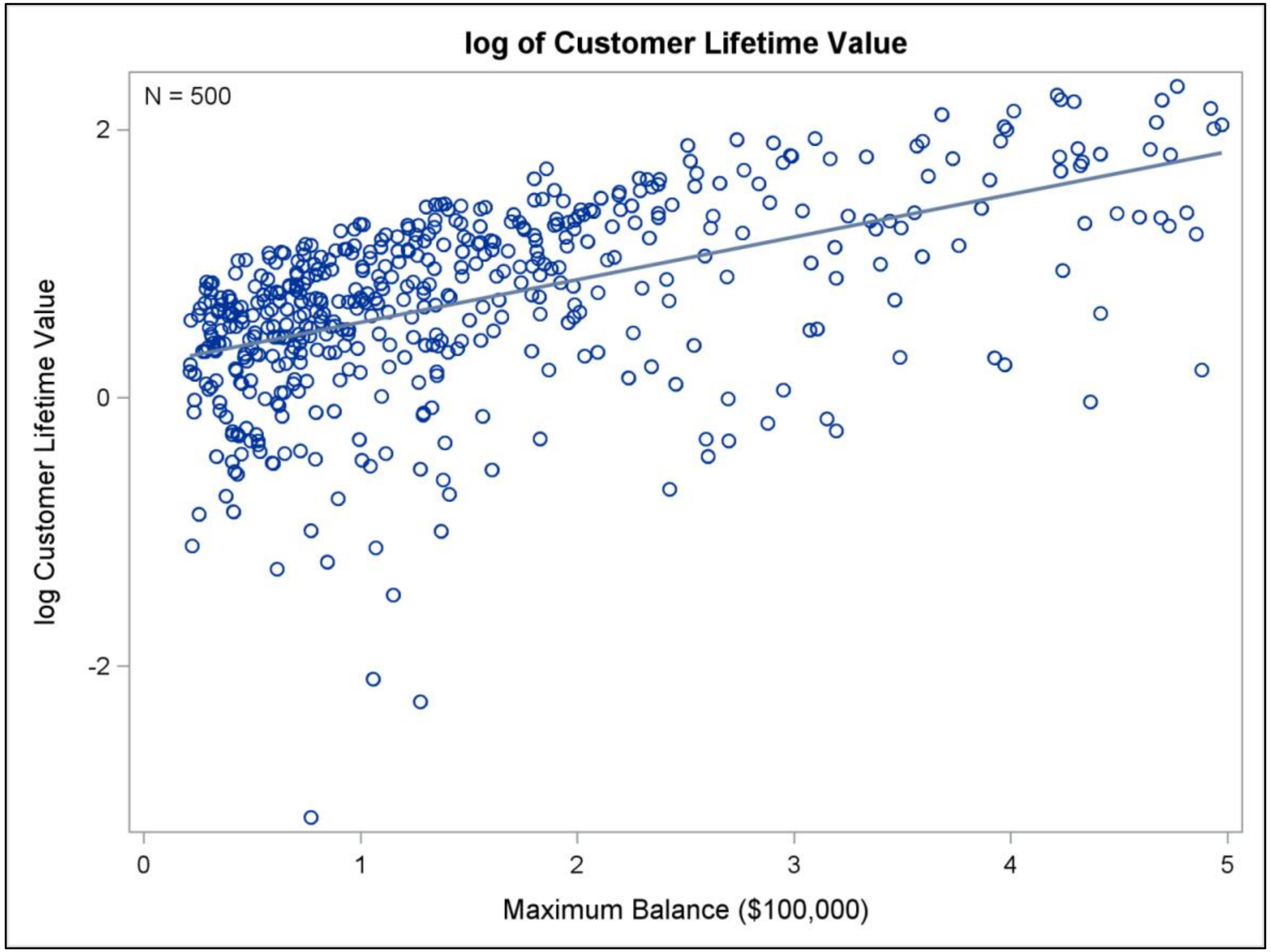
How do you fit a regression model when your data look like this?



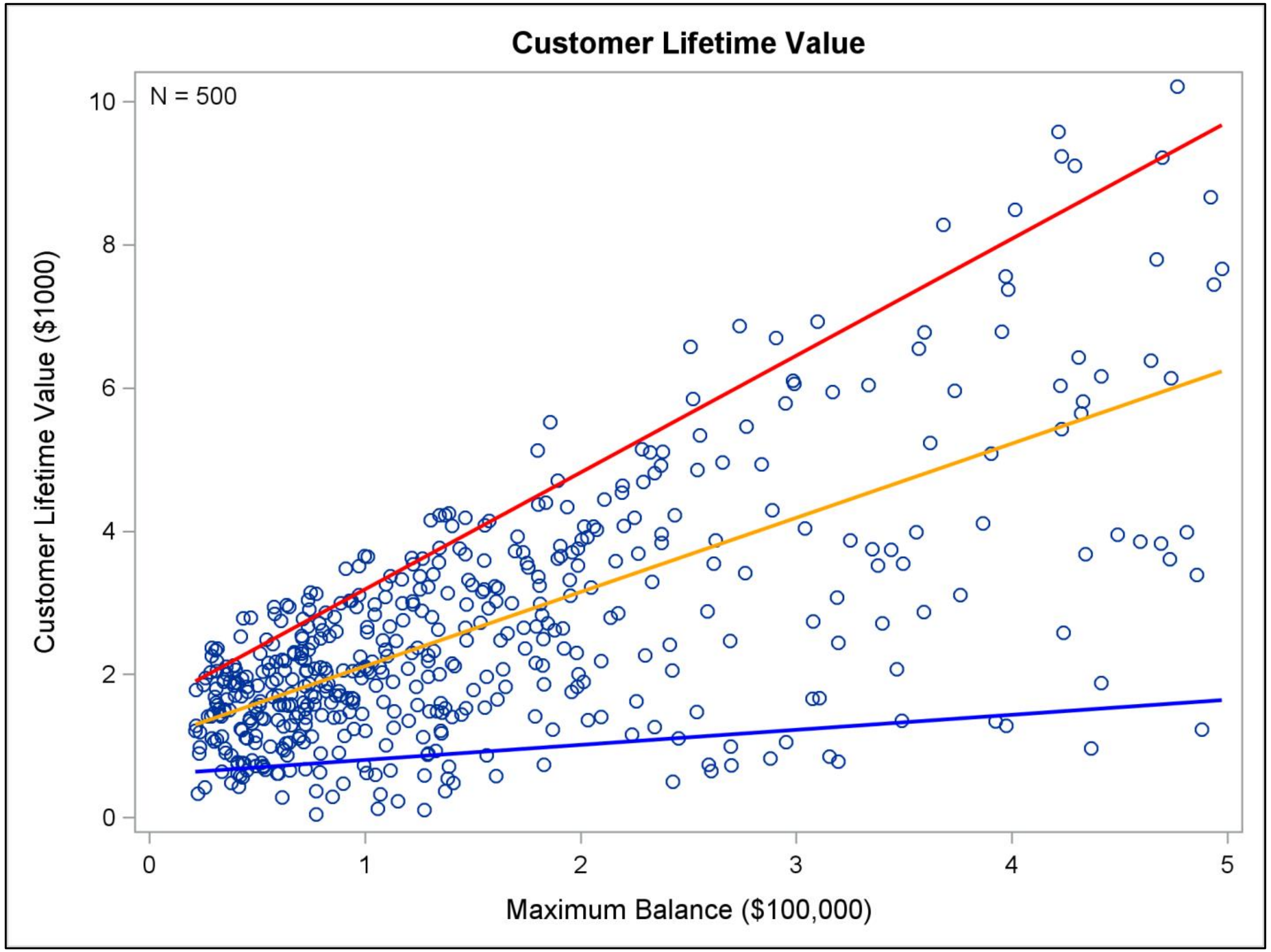
Standard linear regression assumes a constant variance, which is often not the case …



… and applying a preliminary log transformation does not necessarily stabilize the variance



Regression models for percentiles can capture the entire conditional distribution

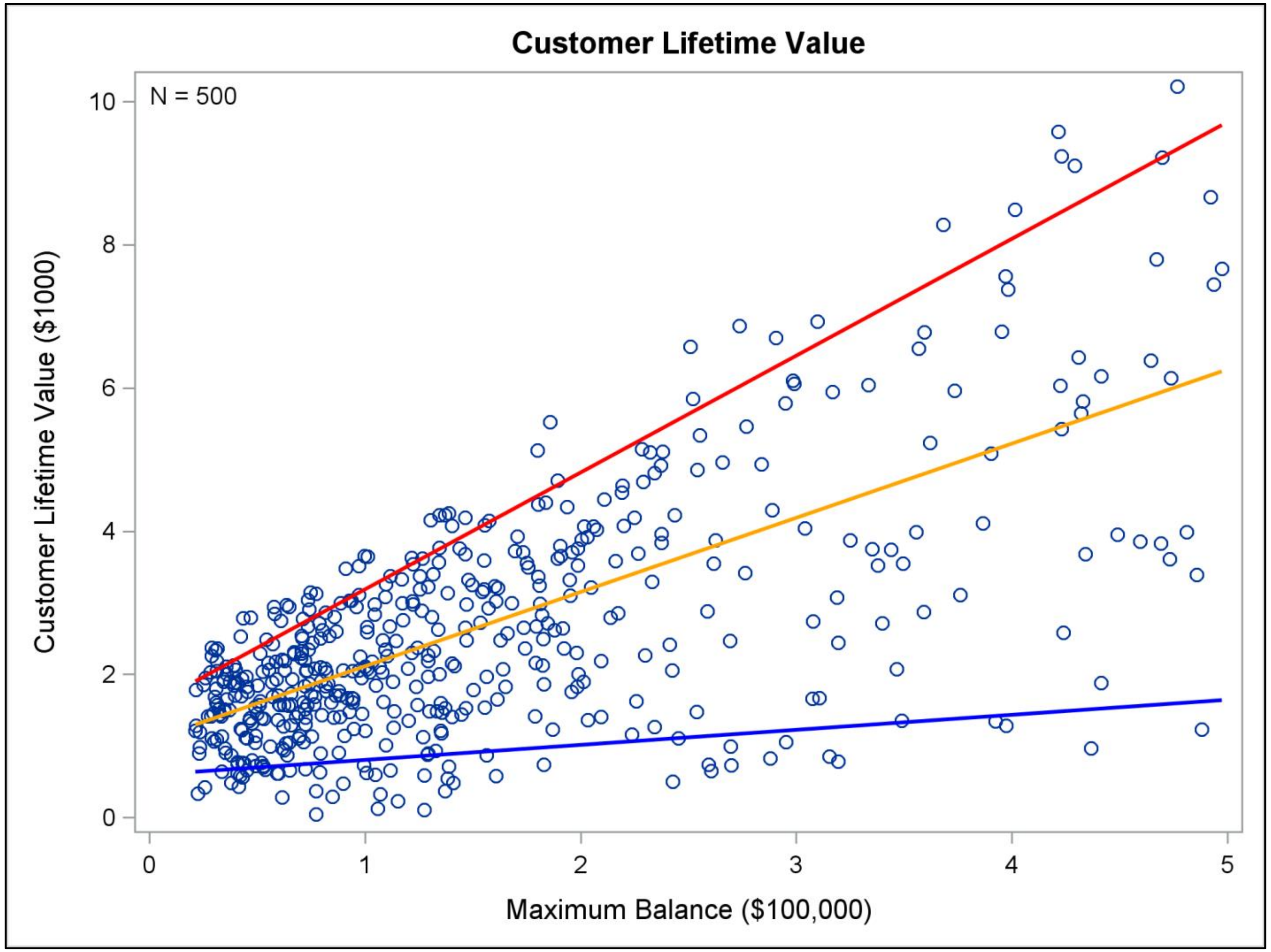


90th percentile

50th percentile

10th percentile

Statisticians use the term quantile in place of percentile, but they have the same meaning …

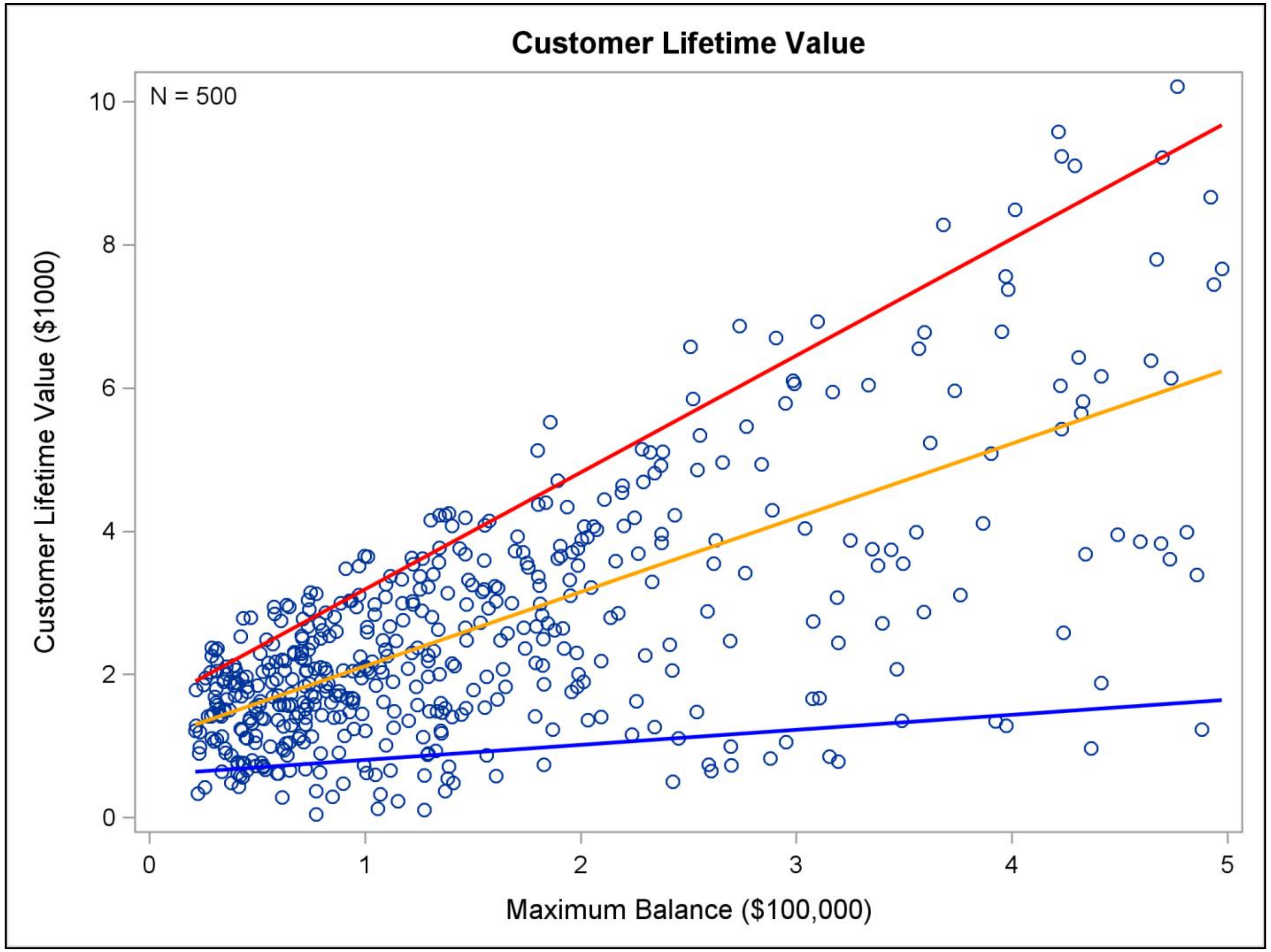


0.9 quantile

0.5 quantile

0.1 quantile

… and the Greek symbol τ denotes the quantile level, which is the probability level associated with the quantile or percentile



90th percentile (τ=0.9)

50th percentile (τ=0.5)

10th percentile (τ=0.1)

How does quantile regression compare with standard linear regression?

Linear Regression	**Quantile Regression**
Predicts conditional mean	Predicts conditional distribution
Applies with limited *n*	Needs sufficient data in tails
Assumes normality	Is distribution agnostic
Is sensitive to outliers	Is robust to outliers
Is computationally inexpensive	Is computationally intensive

Fitting Quantile Regression Models

The coefficient estimates for standard regression minimize a sum of squares

The regression model for the average response is

$$E(y_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} \ , \qquad i = 1, \ldots, n$$

and the β_j's are estimated as

$$\underset{\beta_0, \ldots, \beta_p}{\arg\min} \sum_{i=1}^{n} \left(y_i - (\beta_0 + x_{i1}\beta_1 + \cdots + x_{ip}\beta_p) \right)^2$$

In contrast, the coefficient estimates for quantile regression minimize a sum of "check losses"

The regression model for the τth quantile of the response is

$$Q_\tau(y_i) = \beta_0(\tau) + \beta_1(\tau)x_{i1} + \cdots + \beta_p(\tau)x_{ip} \;, \quad i = 1, \ldots, n$$

and the $\beta_j(\tau)$'s are estimated as

$$\underset{\beta_0, \ldots, \beta_p}{\arg\min} \sum_{i=1}^{n} \rho_\tau \left(y_i - \left(\beta_0 + x_{i1}\beta_1 + \cdots + x_{ip}\beta_p \right) \right)$$

where $\rho_\tau(r) = \tau \max(0, r) + (1 - \tau) \max(0, -r)$

For each level τ, there is a distinct set of regression coefficients

The QUANTREG procedure fits quantile regression models and performs statistical inference

Example

Model the 10th, 50th, and 90th percentiles of customer lifetime value (CLV)

Goal

Target customers with low, medium, and high value after adjusting for 15 covariates, such as maximum balance and average overdraft

```
proc quantreg data=CLV ci=sparsity;
   model CLV = X1-X15 / quantile = 0.1 0.5 0.9;
run;
```

Quantile regression produces a distinct set of parameter estimates and predictions for each quantile level

10th Percentile

Parameter Estimates							
Parameter	DF	Estimate	Standard Error	95% Confidence Limits		t Value	Pr > |t|
Intercept	1	9.9046	0.0477	9.8109	9.9982	207.71	<.0001
X1	1	0.8503	0.0428	0.7662	0.9343	19.87	<.0001
X2	1	0.9471	0.0367	0.8750	1.0193	25.81	<.0001
X3	1	0.9763	0.0397	0.8984	1.0543	24.62	<.0001

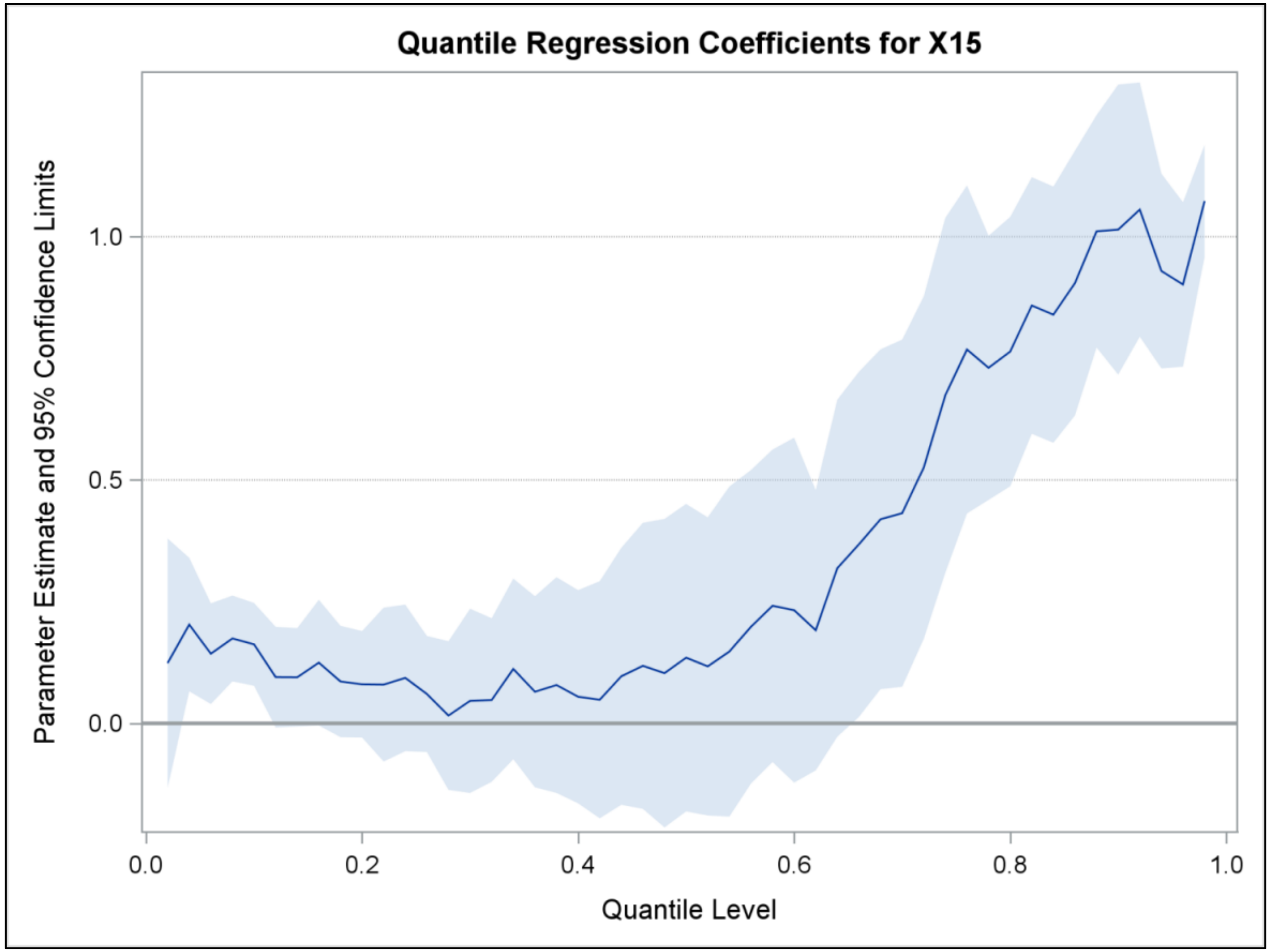
90th Percentile

Parameter Estimates							
Parameter	DF	Estimate	Standard Error	95% Confidence Limits		t Value	Pr > |t|
Intercept	1	10.1007	0.1386	9.8283	10.3730	72.87	<.0001
X1	1	0.0191	0.1485	-0.2726	0.3109	0.13	0.8975
X2	1	0.9539	0.1294	0.6996	1.2081	7.37	<.0001
X3	1	0.0721	0.1328	-0.1889	0.3332	0.54	0.5874

The QUANTREG procedure provides extensive features for statistical inference

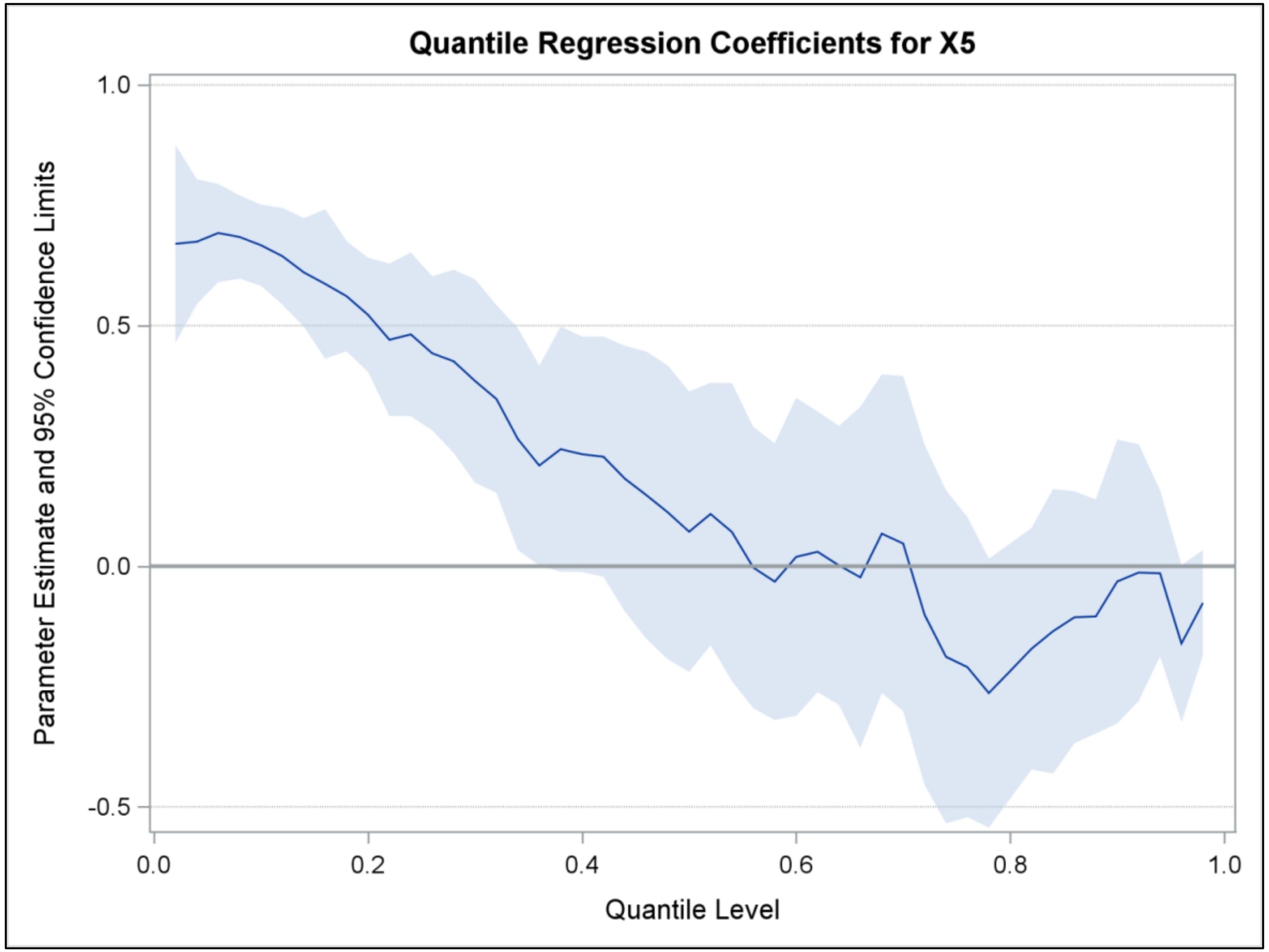
- Simplex, interior point, and smooth algorithms for estimation
- Sparsity and bootstrap resampling methods for confidence limits
- Wald, likelihood ratio, and rank-score tests
- Quantile process regression, which fits a model for all τ in (0,1)

Quantile process plots display the effects of predictors on different parts of the response distribution



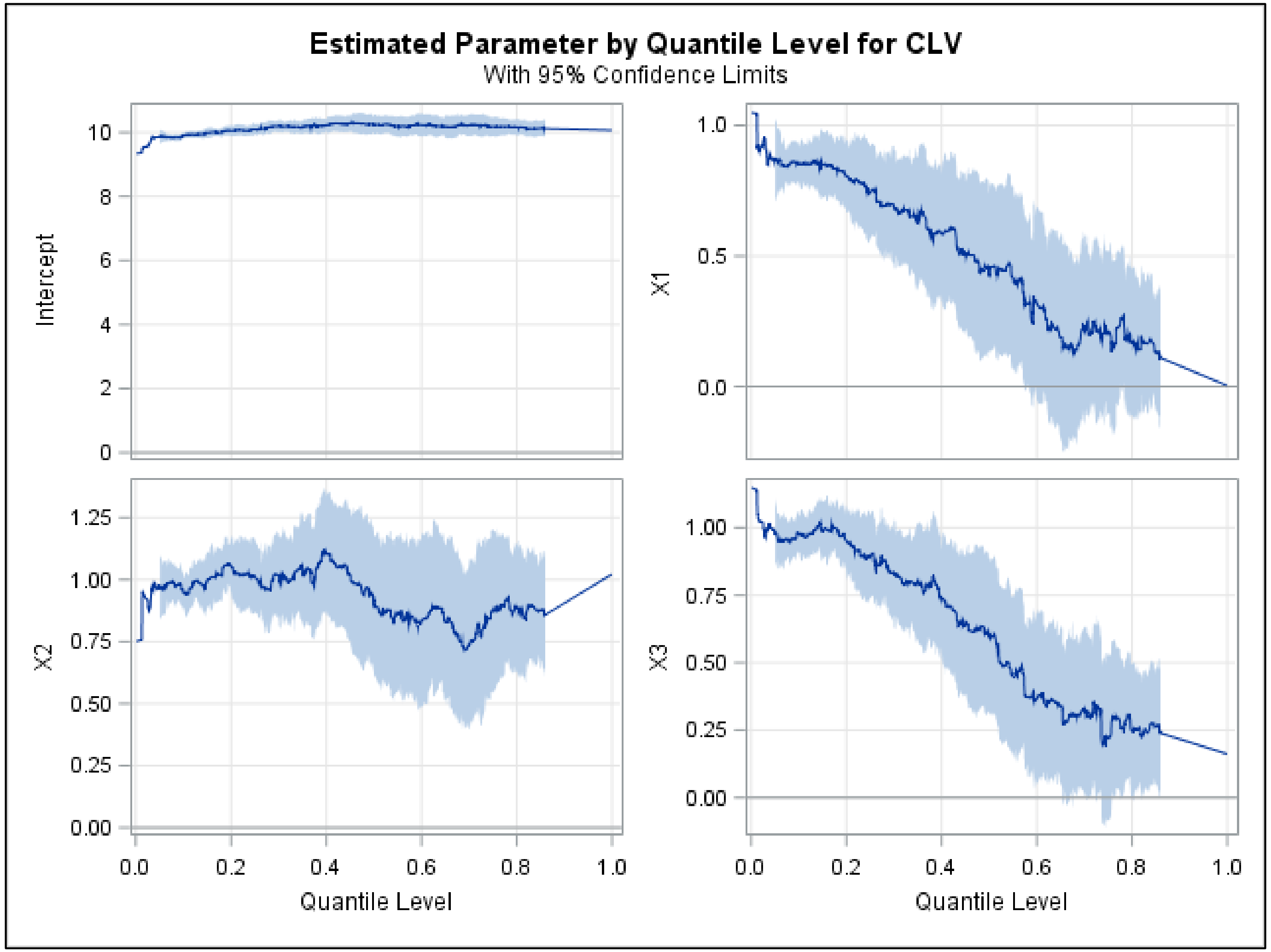
X15 positively affects the upper tail of the distribution

Quantile process plots display the effects of predictors on different parts of the response distribution



X5 positively affects the lower tail of the distribution

Paneled process plots help you identify which predictors are associated with different parts of the response distribution



Building Quantile Regression Models

Example: Which variables differentiate high-performingstores from low-performing stores?

Response: close rates for 500 stores

Candidate predictors

- Store descriptors (X1–X20)
- Promotion (P1–P6)
- Layout (L1–L6)

Approach

1. Build sparse regression models for the 10th, 50th, and 90th percentiles
2. Compare the variables selected for each model

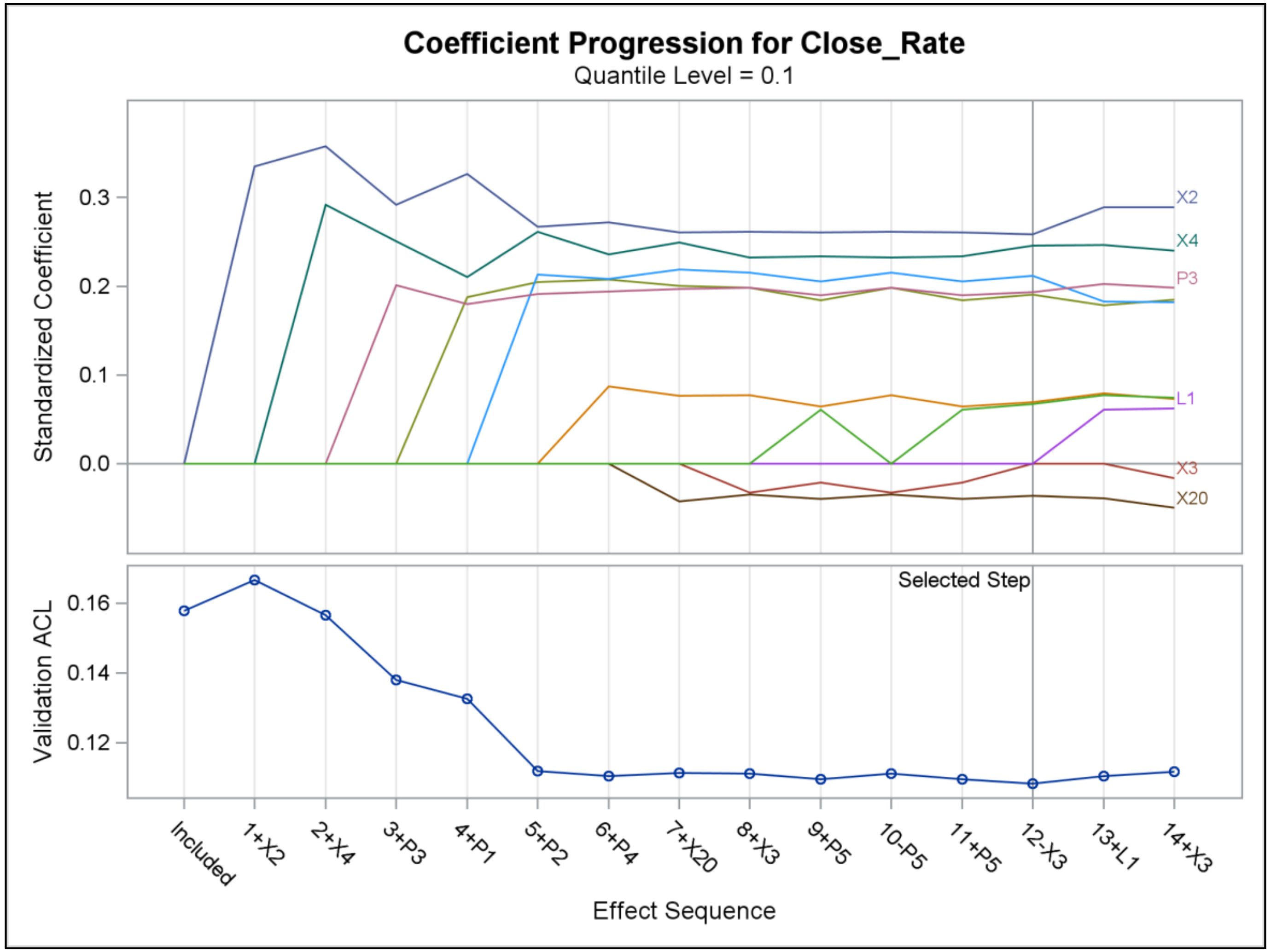
The QUANTSELECT procedure selects effects in quantile regression models

Features

- Provides forward, backward, stepwise, and lasso selection methods
- Provides extensive control over the selection
- Builds models for specified quantiles or the entire quantile process

```
proc quantselect data=Store plots=Coefficients;
   model Close_Rate = X1-X20 L1-L6 P1-P6 /
         quantile=0.1 0.5 0.9  selection=lasso(sh=3);
   partition fraction(validate=0.3);
run;
```

Coefficient progression plots show how the model fit evolves during variable selection



The layout variables L2, L3, and L5 are selected only in the model for the 90th percentile of close rates

10th Percentile

Parameter Estimates			
Parameter	DF	Estimate	Standardized Estimate
Intercept	1	60.097618	0
X2	1	0.953402	0.258498
X4	1	0.933705	0.245902
X20	1	-0.140895	-0.035981
P1	1	0.724145	0.190798
P2	1	0.783880	0.211752
P3	1	0.696274	0.193163
P4	1	0.260641	0.069442
P5	1	0.242147	0.067135

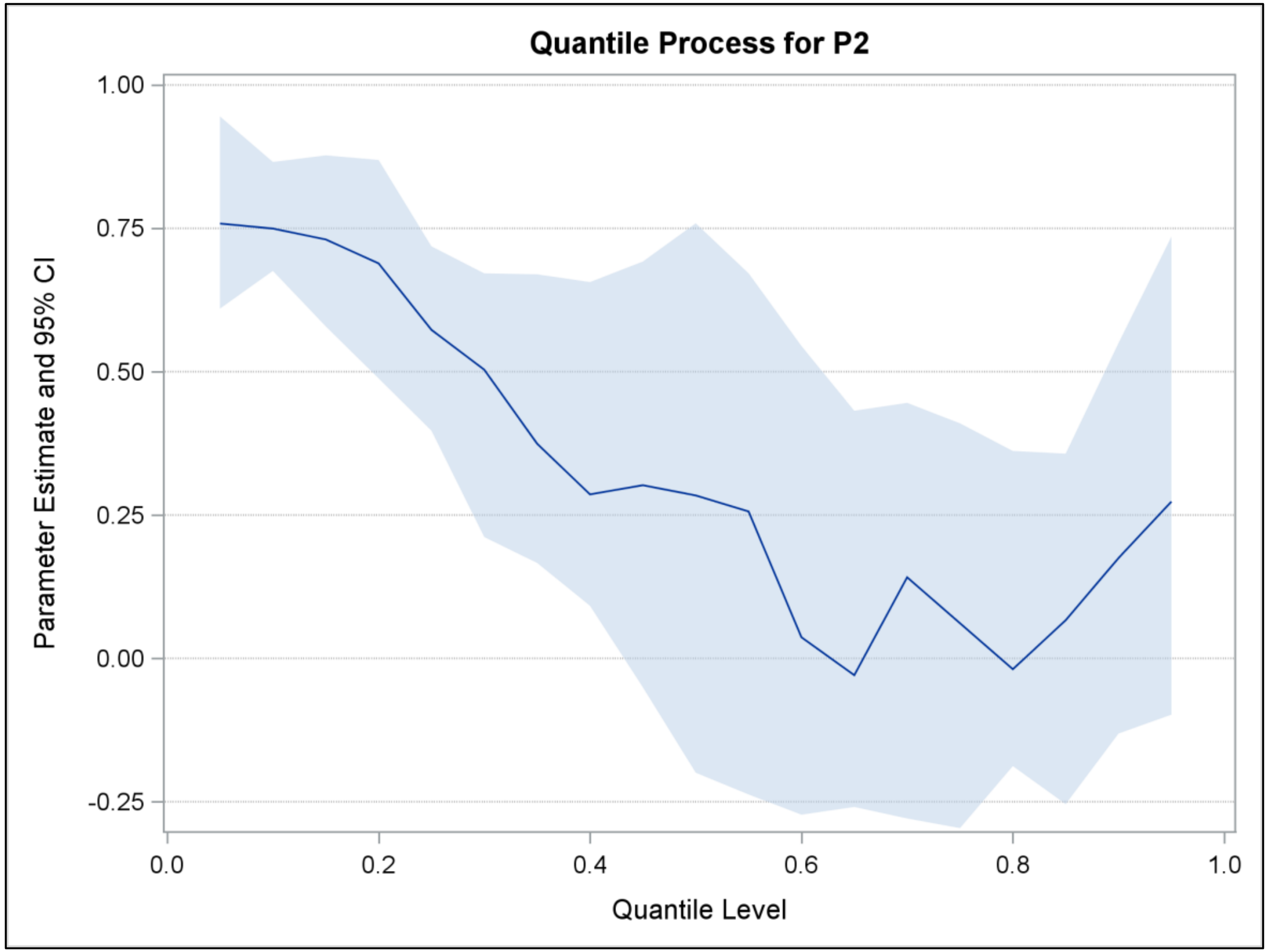
50th Percentile

Parameter Estimates			
Parameter	DF	Estimate	Standardized Estimate
Intercept	1	60.950579	0
X2	1	1.508595	0.409029
X4	1	0.710687	0.187168
P3	1	0.361047	0.100163
P4	1	0.669943	0.178491
P5	1	0.544278	0.150902

90th Percentile

Parameter Estimates			
Parameter	DF	Estimate	Standardized Estimate
Intercept	1	61.079231	0
X2	1	0.982776	0.266463
X4	1	1.118507	0.294572
L2	1	1.027725	0.297930
L3	1	0.859988	0.240257
L5	1	0.672210	0.186588
P5	1	0.192967	0.053500

Quantile regression gives you insights that would be difficult to obtain with standard regression methods



P2 positively affects the lower half of the close rate distribution

The syntax and features of the QUANTSELECT procedure are similar to those of the GLMSELECT procedure

- Models can contain main effects consisting of continuous and classification variables, and their interactions
- Models can contain constructed effects, such as splines
- Each level of a CLASS variable can be treated as an individual effect
- Data can be partitioned to avoid overfitting

Application to Risk Management

Quantile regression provides a robust approach for estimating value at risk (VaR)

- VaR measures market risk by how much a portfolio can lose within a given time period, for a confidence level $(1 - \tau)$

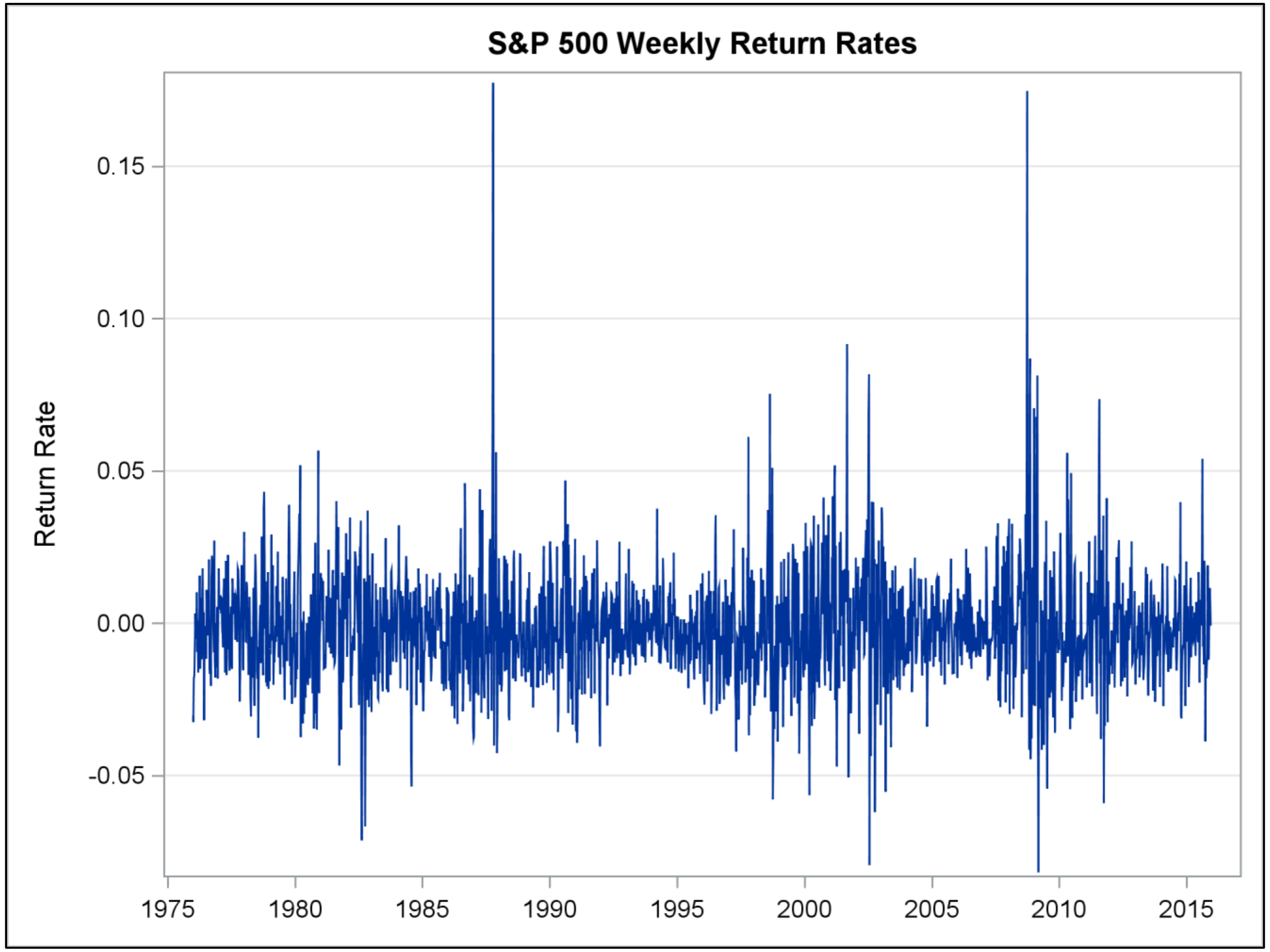
- VaR is a conditional quantile of future portfolio values

$$\Pr[\, y_t < -\mathrm{VaR}_t \mid \Omega_t \,] = \tau$$

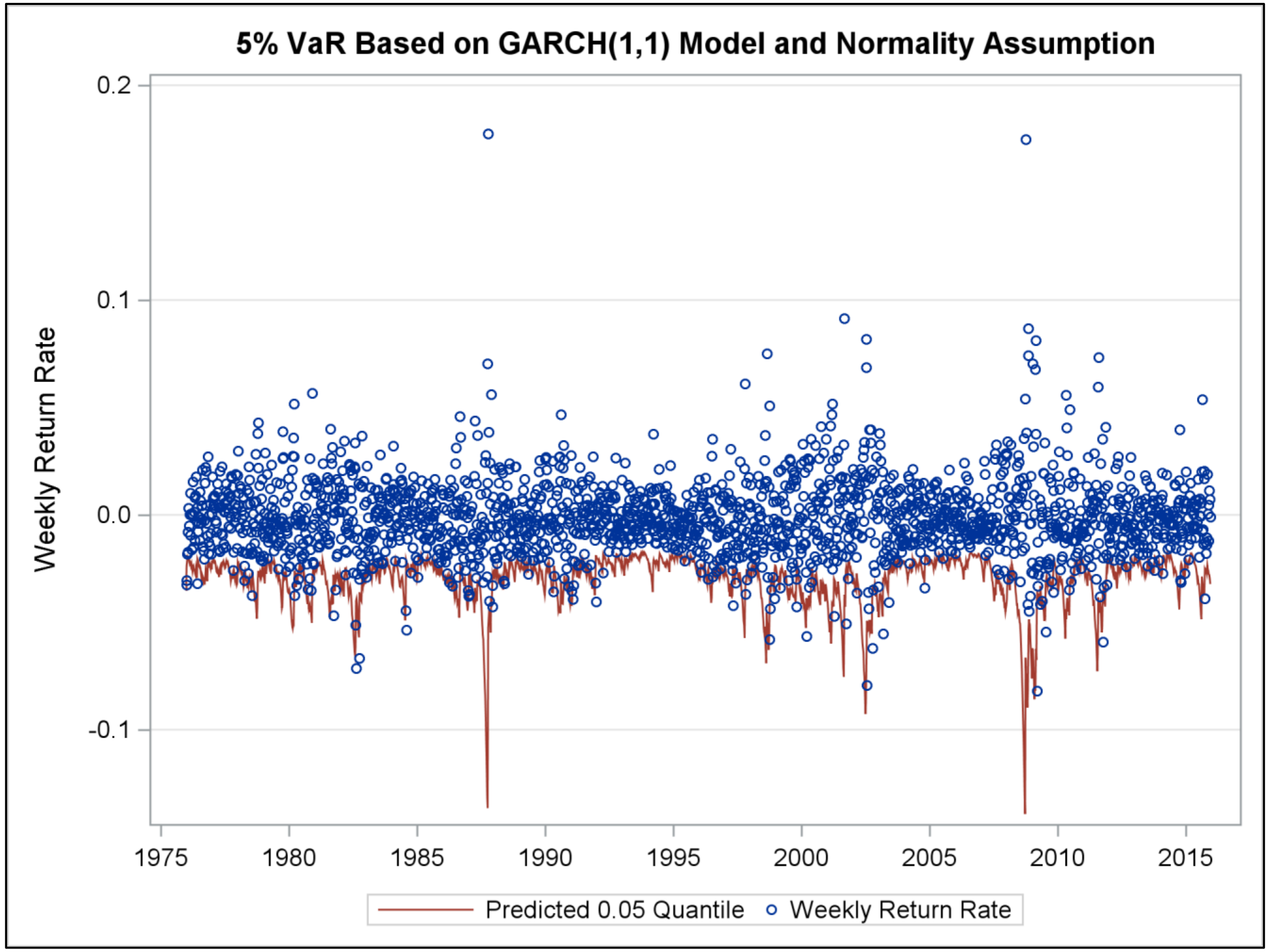
where Ω_t is the information at time t and $\{y_t\}$ is the series of financial returns

- Methods of measuring VaR include GARCH models, which estimate the volatility of the portfolio and assume the returns are normally distributed

GARCH models have been applied to the weekly return rates of the S&P 500 Index, which display skewness and heavy tails



You can use PROC VARMAX to predict VaR with a GARCH(1,1) model, which assumes normality …



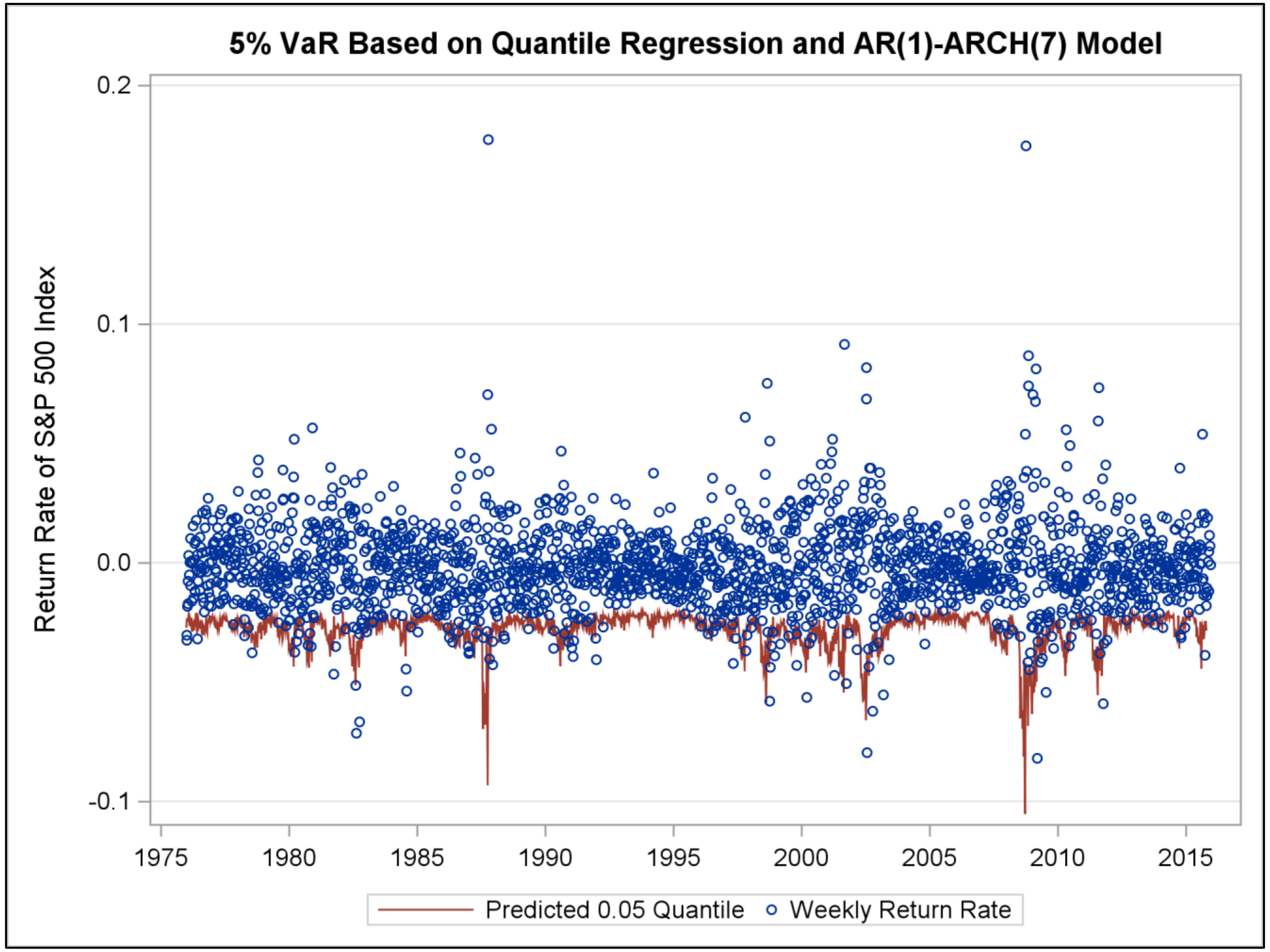
… or you can use PROC QUANTREG to predict VaR by conditioning on lagged standard errors estimated by PROC VARMAX

```
proc varmax data=SP500;
   model Rate / p=1;
   garch form=ccc subform=garch q=6;
   output out=StdErr lead=1;
   id date interval=week;
run;

proc quantreg data=StdErr;
   model Rate = std1-std7 / quantile=0.05;
   output out=qr p=VaR;
   id date;
run;
```

Xiao, Guo, and Lam (2015)

Quantile regression offers robustness in situations where market returns display negative skewness and excess kurtosis



Xiao, Guo, and Lam (2015)

Application to Ranking Student Exam Performance

How would you rank two students, Mary and Michael, who took the same college entrance exam?

- Mary scored 1948 points, and her quantile level is

$$\Pr[\text{ Score} \leq 1948 \text{ }] = 0.9$$

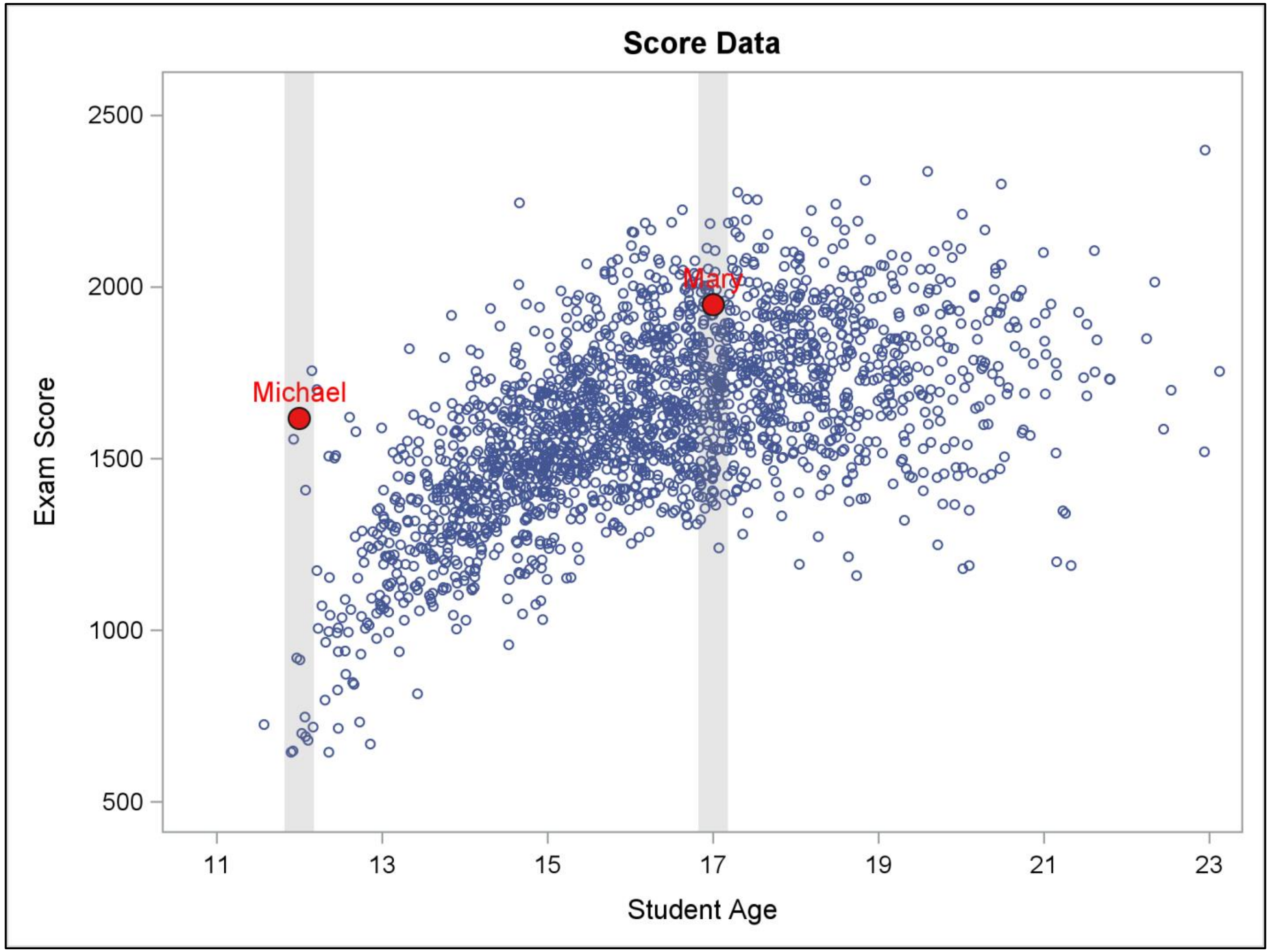
- Michael scored 1617 points, and his quantile level is

$$\Pr[\text{ Score} \leq 1617 \text{ }] = 0.5$$

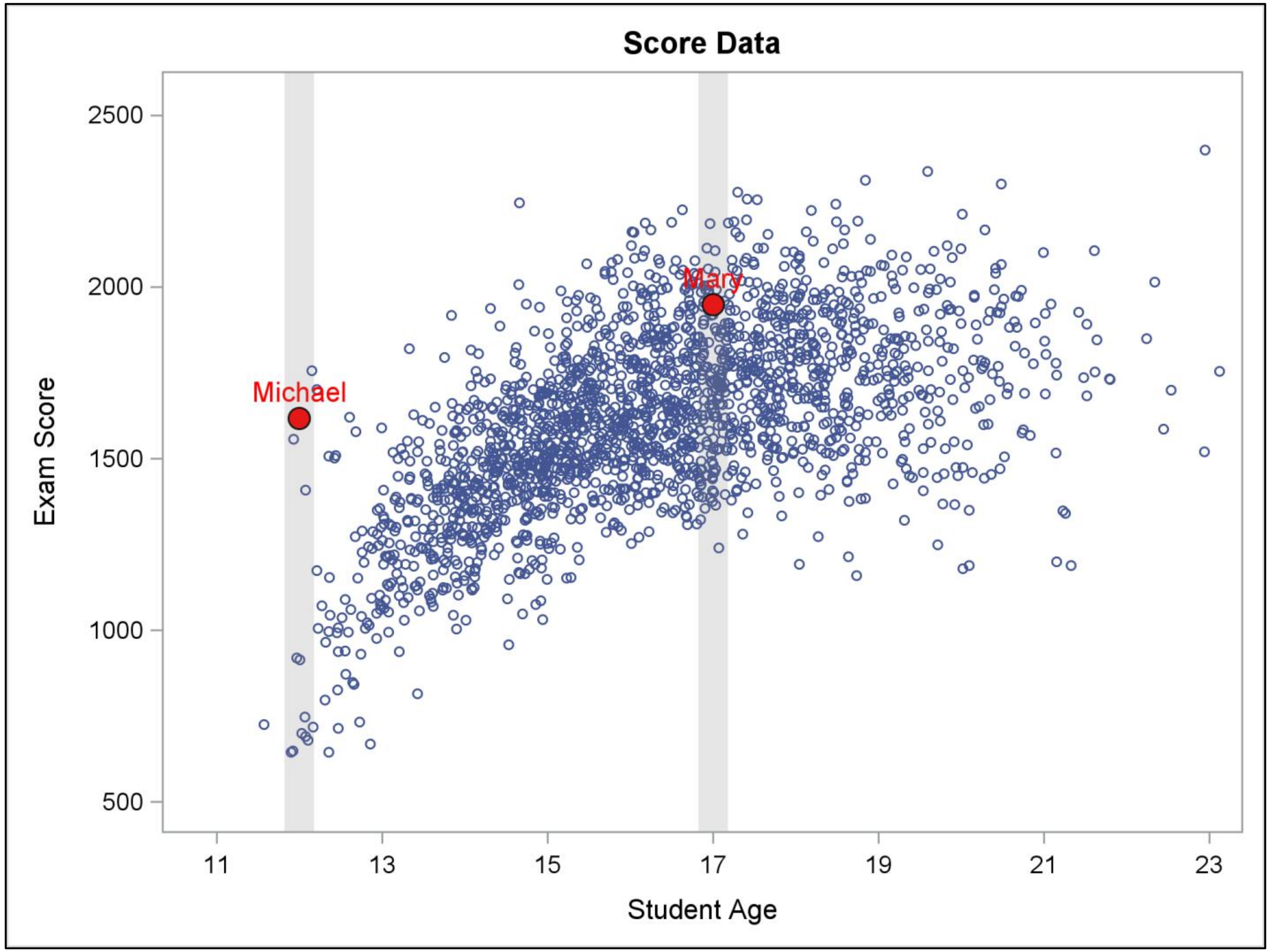
- Now you learn that Mary is age 17 and Michael is age 12
- To rank them, you need to determine their *conditional* quantile levels

$$\Pr[\text{ Score} \leq 1948 \mid \text{Age} = 17]$$
$$\Pr[\text{ Score} \leq 1617 \mid \text{Age} = 12]$$

Where do Michael and Mary fall within the score distributions for their age groups?



What are Michael's and Mary's quantile levels based on the score distributions for their age groups?

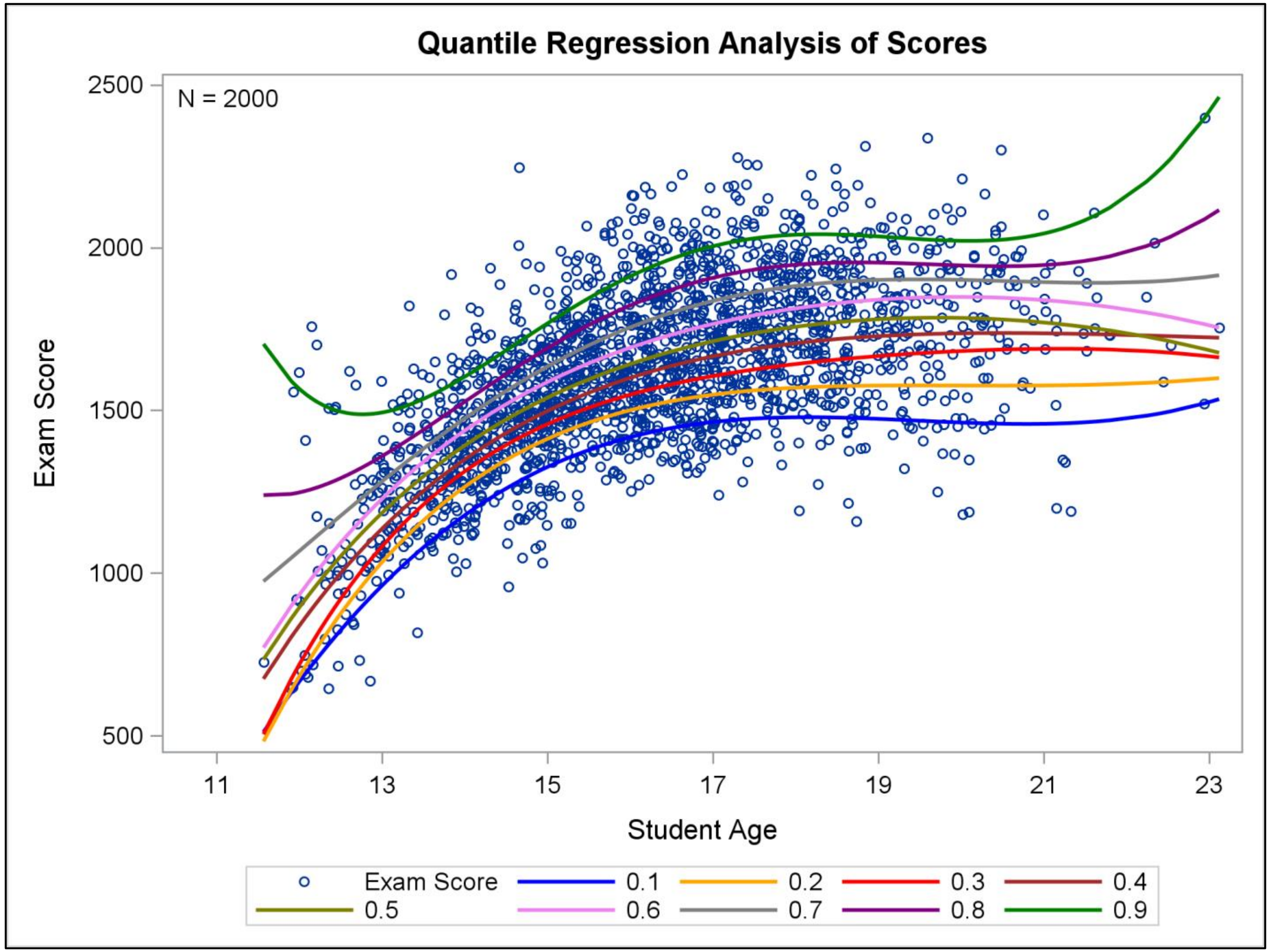


You can estimate the conditional distributions by using quantile regression

1. Use PROC QUANTREG to fit a quantile regression model that predicts the quantiles for an extensive grid of levels, such as 0.01, 0.02, …, 0.99
2. From the quantiles, estimate the conditional distributions of the response for covariate values corresponding to specified observations
3. Compute the predicted quantile (percentile) *levels* from the distributions, and use these to rank the observations

The QPRFIT macro, new in SAS/STAT® 14.2, implements all three steps

Begin by modeling the conditional quantiles of Score for a uniform grid of quantile levels

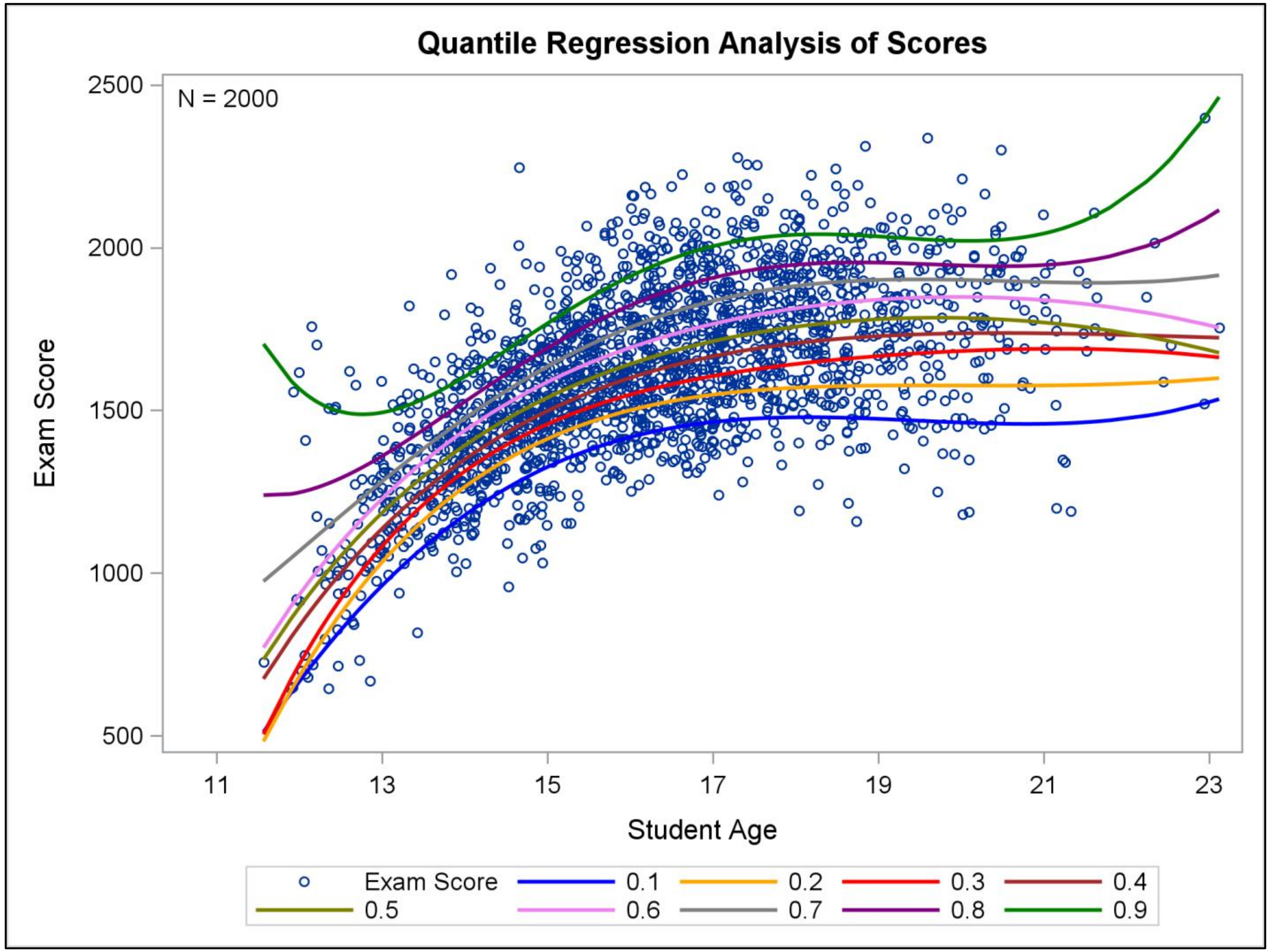


It is important to specify an appropriate model with terms that capture the nonlinearity in the data

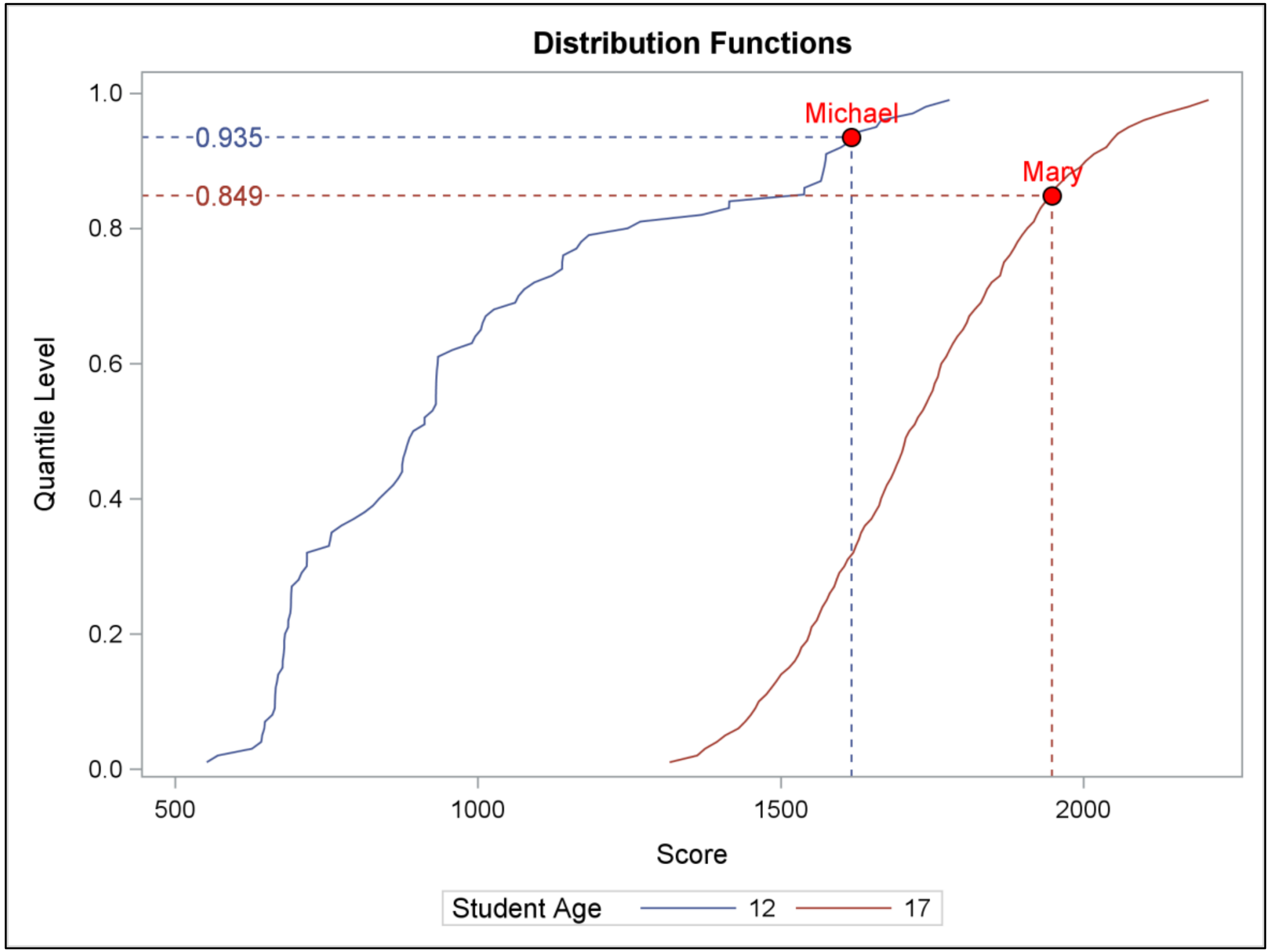
```
data Score;
   set Score;
   Age2   = Age*Age;
   Age3   = Age2*Age;
   AgeInv = 1/Age;
run;

proc quantreg data=Score;
   model Score = Age Age2 Age3 AgeInv /
                 quantile = 0.10 to 0.90 by 0.1;
   output out=ModelFit p=Predicted;
run;
```

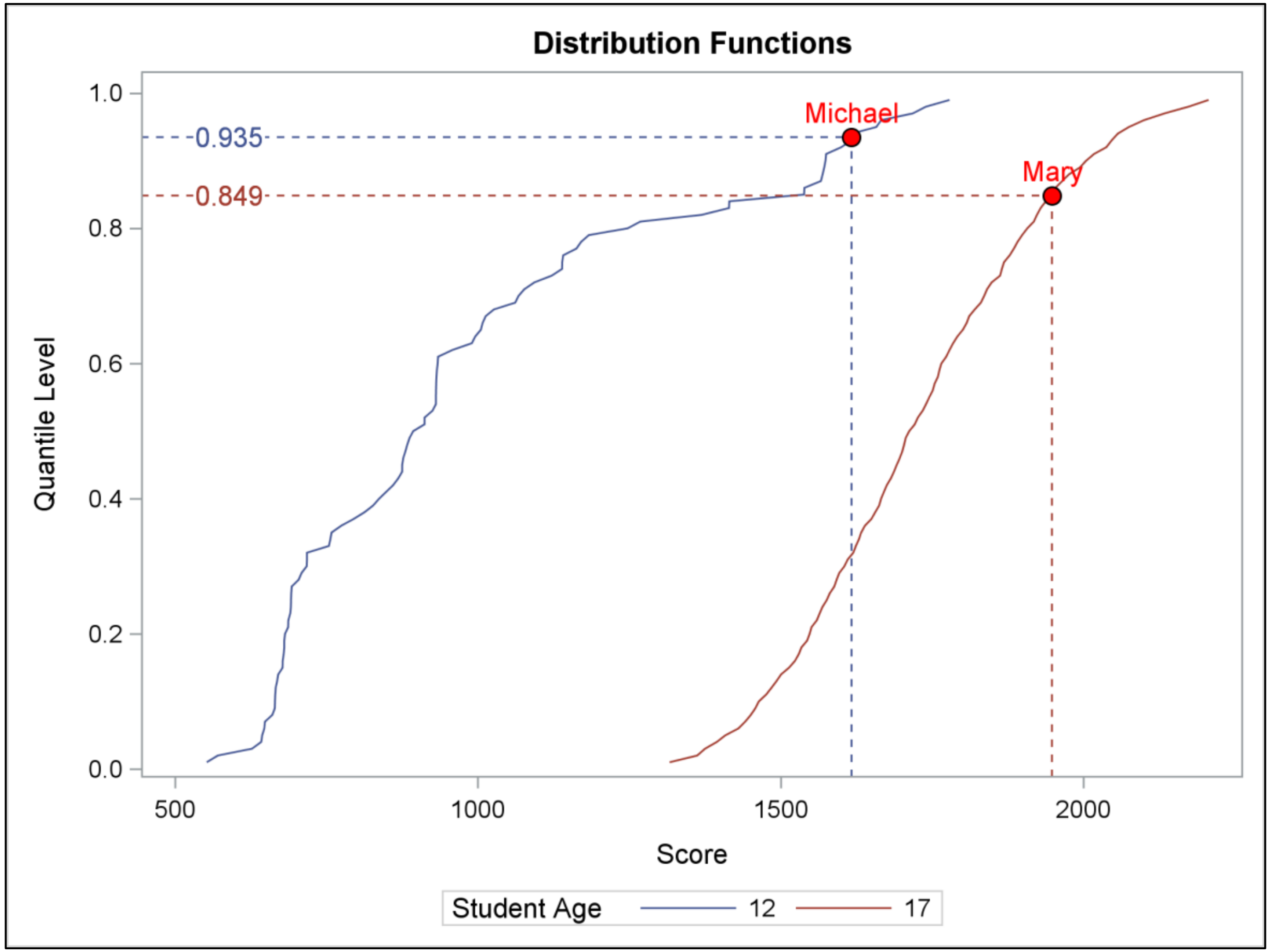
Note that the shape of the conditional distribution for Score differs with Age



The QPRFIT macro uses the predicted quantiles to compute the conditional distribution functions of Score for Age=12 and Age=17



Evaluating the conditional distributions at the scores for Michael and Mary provides their adjusted quantile levels



How do Michael and Mary rank before and after adjusting for their ages?

Obs	Name	Score	Age	Mean	Median	Regression Quantile Level	Sample Quantile Level
1	Michael	1617	12	971.43	893.45	0.93500	0.50075
2	Mary	1948	17	1709.94	1712.36	0.84851	0.90025

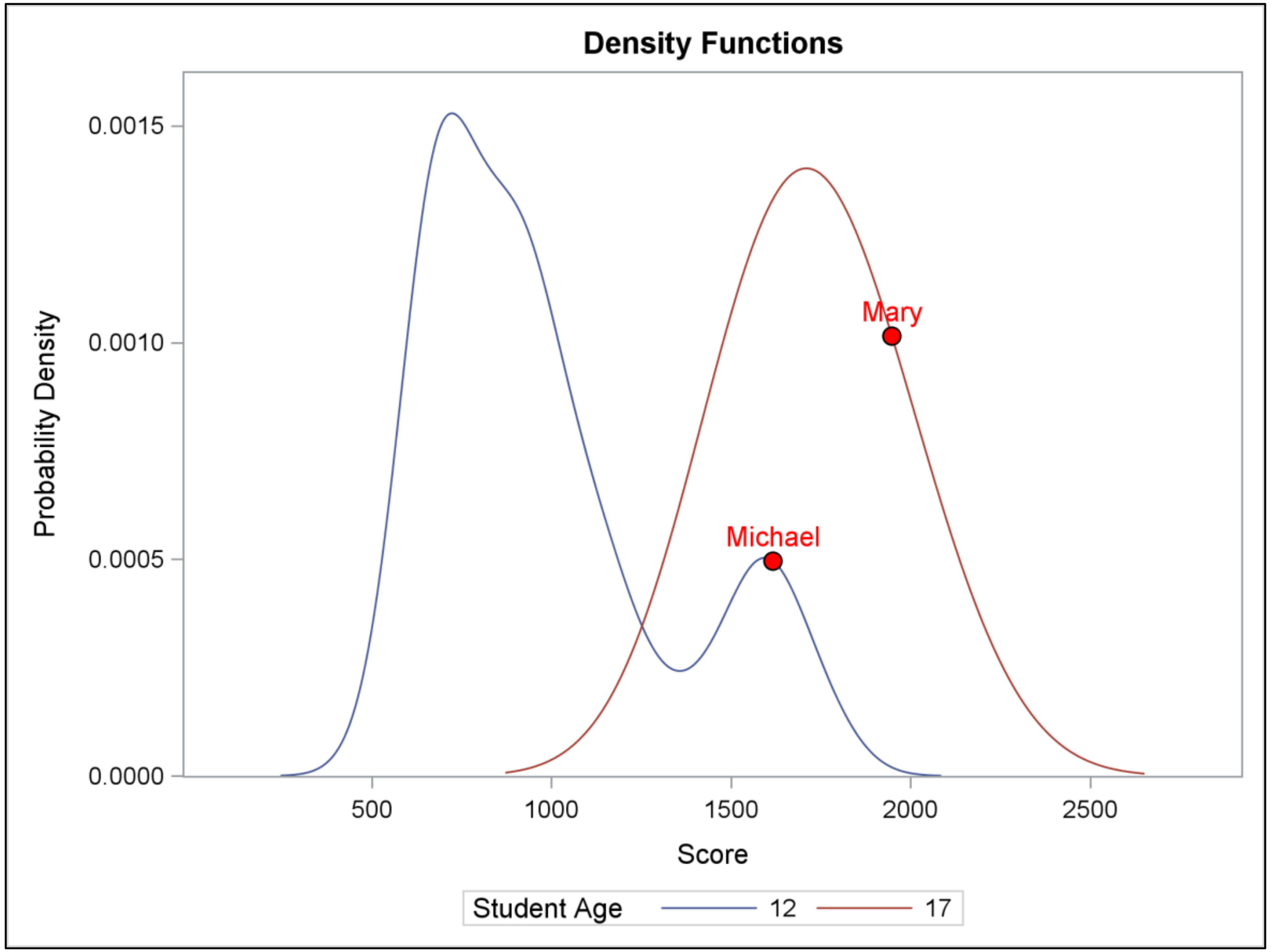
The QPRFIT macro fits a quantile regression model and computes adjusted quantiles for specified observations

```
data ScoreID;
   Name='Michael'; output;
   Name='Mary';    output;
run;

%qprFit(data=Score, depvar=Score,
        indvar=Age Age2 Age3 AgeInv, onevar=Age,
        nodes=99, iddata=ScoreID,
        showPDFs=1, showdist=1)
```

The proceedings paper explains how to use the macro

The QPRFIT macro also estimates the probability density functions for Age=12 and Age=17



Wrap-Up

Five points to remember for using quantile regression in your work

1. Quantile regression is versatile because it allows a general linear model and does not assume a parametric distribution

2. Quantile regression estimates the entire conditional distribution and allows its shape to depend on predictors

3. Quantile process plots reveal effects of predictors on different parts of the response distribution

4. Quantile regression can predict quantile levels of observations while adjusting for effects of covariates

5. The QUANTREG and QUANTSELECT procedures are powerful tools for fitting and building models, even with large data

Learn more at http://support.sas.com/statistics

Welcome to Statistics and Operations Research

SAS has long developed software for data analysis, econometrics, operations research, and quality improvement. The purpose of these pages is to provide our users with technical information about using this software, including details about software capabilities, examples, papers, e-newsletter, and communities.

SAS/ETS | SAS/IML | SAS/OR | SAS/QC | SAS/STAT

SAS/ETS software offers a broad array of time series, forecasting, and econometric techniques that enable modeling, forecasting, and simulation of business processes for improved strategic and tactical planning.

SAS/ETS 14.2 introduces the new SASENOAA and SASERAIN interface engines as well as the new SPATIALREG procedure along with new features for the HPCDM, HPSEVERITY, SEVERITY, QLIM, SSM, TIMESERIES, and VARMAX procedures and the SASEFAME, SASEFRED, SASEQUAN, and SASEXFSD interface engines.

Read about What's New in SAS/ETS 14.2.

Featured News

SAS Global Forum 2017

If you are still debating whether to attend this year, consider what you might miss:

- Statistical Tutorials on Sunday
- Over a dozen presentations by SAS staff on new capabilities
- Numerous emerging technology presentations

Upcoing Conferences

2017 American Statistical Association Conference on Statistical Practice
Feb. 23–25, 2017
Jacksonville, FL

ENAR 2017
Mar. 12–15, 2017
Washington, DC

SAS Global Forum 2017
Apr. 2–5, 2017
Orlando, FL

HIGHLIGHTS

- Latest e-Newsletter
- 14.2 Highlights and Information
- SAS Analytical Handouts
- SAS Macros for Experimental Design and Choice Modeling
- ODS Statistical Graphics

VIDEOS

- Computing an Optimal Blackjack Strategy with SAS/OR New
- SAS Simulation Studio: Power Up by Calling SAS Programs New
- What's New in Econometric Modeling in SAS/ETS 14.1
- Spatial Dependence, Nonlinear Panel Models, and More New Features in SAS/ETS 14.1
- Exploring Econometric Tasks in SAS Studio
- Statistical Quality Improvement with SAS/QC
- See all videos

PAPERS

- SAS/STAT 14.1: Methods for Massive, Missing, or Multifaceted Data
- Practical Applications of SAS Simulation Studio

Sign up for e-newsletter

Watch short videos

Download overview papers

Five Things You Should Know about Quantile Regression