NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Quantile Regression

SESSION 4 IN-CLASS SLIDES

Intended Learning Outcomes

Identify aspects of business problems that cause standard analytics models to become useless or less effective.

Apply advanced techniques to overcome or mitigate the weaknesses of standard analytics models.

Evaluate performance of the advanced predictive techniques.

Explain the workings and results of the advanced predictive techniques in the context of the business problem to client/employer.

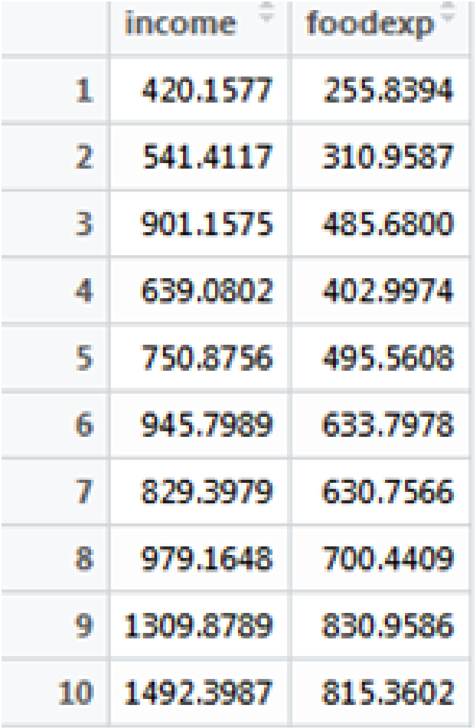
Propose business solutions/recommendations based on the advanced predictive techniques.

Quiz

Ungraded. Check your understanding of this Session Content.
Use your real name (not nickname) in the quiz.

Engel Dataset from Rpackage quantreg

- Dataset that records Family Expenditure on Food and Family Income in Belgium 1857.
- Used to show a limitation of Linear Regression and usefulness of Quantile Regression.
- Dataset is in quantreg Rpackage.



	income	foodexp
1	420.1577	255.8394
2	541.4117	310.9587
3	901.1575	485.6800
4	639.0802	402.9974
5	750.8756	495.5608
6	945.7989	633.7978
7	829.3979	630.7566
8	979.1648	700.4409
9	1309.8789	830.9586
10	1492.3987	815.3602

First 10 of 235 records in Engel Dataset.

Activity 1

Quantile Regression in R

Part of Pre-Class Learning Activity

Est. Duration: 20 mins

1. Run the RScript qr1.R
2. In the RScript qr1, the 6 quantile regression are plotted as 6 grey lines, but the model coefficients (b_0, b_1) are not shown. Modify the RScript so that the model parameters for the 6 quantile regression models are exhibited in a table. [Hint: Where is the information saved in the R object?]
3. One student asked if quantile regression is just fitting linear regression on the specific percentile of the data. True/False? Can you answer this from the software output?

Instructor solution qr2.R will be posted in main site by end of week.

Quantile Regression Model Coefficients at various Percentiles

Quantile Regression Model

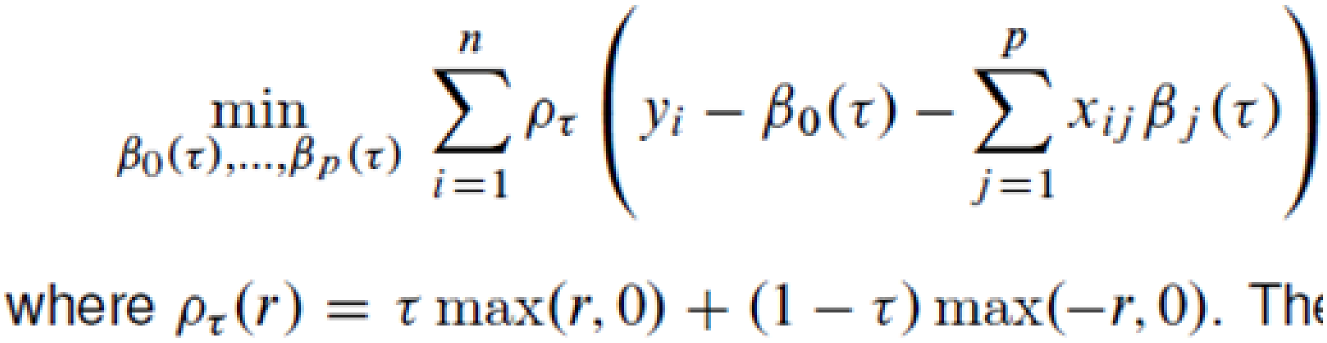
- Uses all the data regardless of quantile.
- Quantile Regression Model:

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

Source: Rodriguez and Yao (2017) Five Things You Should Know about Quantile Regression. Paper SAS525-2017, SAS Institute.

Quantile Regression Model Coefficients

- Linear Reg Model Coefficients (b_0, b_1,…b_p) estimated by minimizing the sum of squared errors.
- Quantile Reg Model Coefficients estimated from minimizing the sum of check function:



$$\min_{\beta_0(\tau),\ldots,\beta_p(\tau)} \sum_{i=1}^{n} \rho_\tau \left(y_i - \beta_0(\tau) - \sum_{j=1}^{p} x_{ij}\beta_j(\tau) \right)$$

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

Source: Rodriguez and Yao (2017) Five Things You Should Know about Quantile Regression. Paper SAS525-2017, SAS Institute.

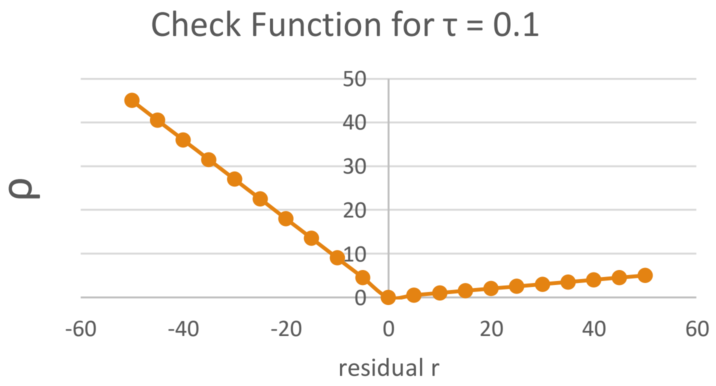
Understanding the Check Function

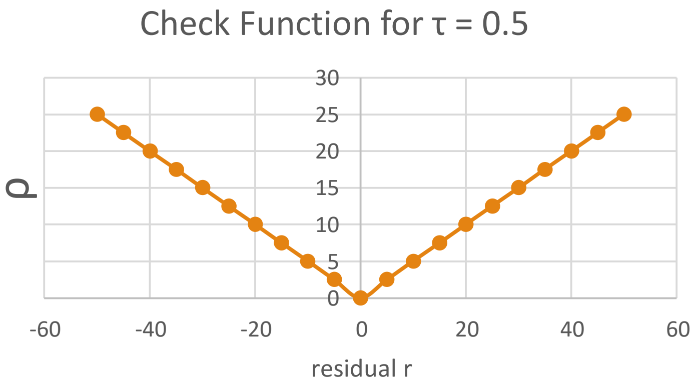
USED IN QUANTILE REGRESSION

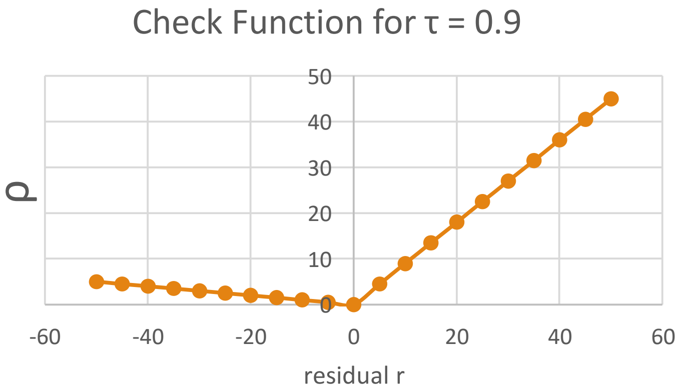
Understanding the Check Function:

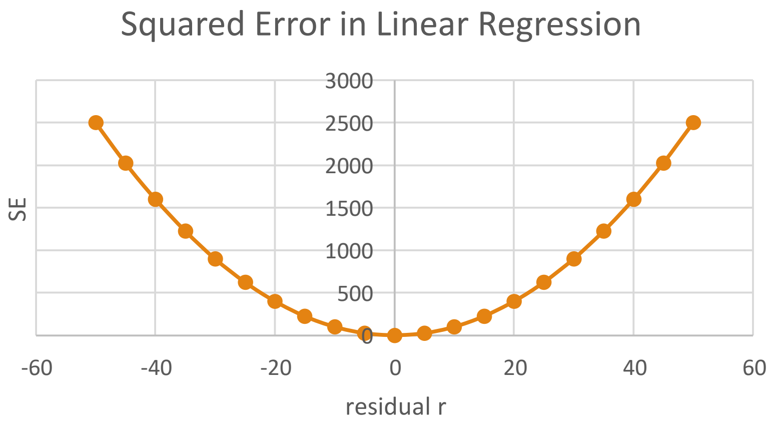
$$\rho_\tau(r) = \tau \max(r, 0) + (1 - \tau) \max(-r, 0)$$

- See Excel File: Check Function > By Tau, for examples using 3 values of τ.
- What can you conclude from the numerical examples?


Check Function for τ = 0.1


Check Function for τ = 0.5


Check Function for τ = 0.9


Squared Error in Linear Regression

Source: Chew C.H. (2020) AI, Analytics and Data Science Vol. 2.

Insights about Check Function

- Different Tau use different “punishment” levels as the “stick” to incentivize the model to predict at the right levels of the y data.
  - The higher the Tau value, the higher the quantile regression function, and hence the model predicted Y value.
- If Tau < 0.5, the punishment is higher if residual < 0. i.e. actual value of Y < model predicted value of Y.
- If Tau = 0.5, the quantile regression model is equally punished if it is lower or higher than the centre.
- If Tau > 0.5, the punishment is higher if residual > 0. i.e. actual value of Y > model predicted value of Y.

Source: Chew C.H. (2020) AI, Analytics and Data Science Vol. 2.

Class Activity 2

Total Loss in 3 Models for Each Tau

Est. Duration: 20 mins

- Open the Excel File: Check Function > Total Loss worksheet.
- Given 4 data points, 3 models Yhat1, Yhat2, and Yhat3, and 3 values of Tau (0.1, 0.5, 0.9), fill up the blue and yellow cells in Excel.

1. For each model, which tau value result in the lowest total loss?
2. Did we use all the given data points to compute total loss, regardless of the value of tau? Yes/No.
3. What is your conclusion about the tau value and height of the quantile regression line?

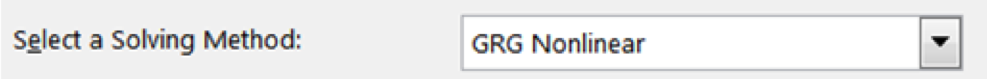
Source: Chew C.H. (2020) AI, Analytics and Data Science Vol. 2.

Ans to Class Activity 2

Source: Chew C.H. (2020) AI, Analytics and Data Science Vol. 2.

Demo: Using Excel Solver to Minimize Total Loss for Linear Regression

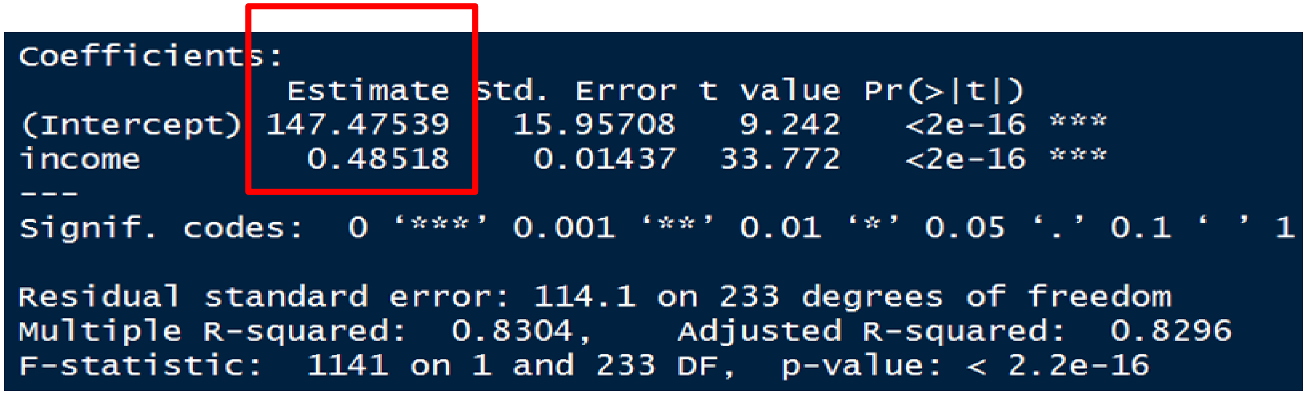
- Plan:
  - Define the Total Loss metric to be minimized.
  - Use Excel Solver to find the values for the linear regression model coefficients b_0 and b_1 that will minimize the Total Loss.
  - Select solving method to GRG Nonlinear



Select a Solving Method: GRG Nonlinear

Note: If Solver is missing in Excel, activate it via File > Options > Add-Ins.

- Check Solver results. Are the answers the same as R Linear Regression results?



```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 147.47539   15.95708   9.242   <2e-16 ***
income        0.48518    0.01437  33.772   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 114.1 on 233 degrees of freedom
Multiple R-squared:  0.8304,	Adjusted R-squared:  0.8296 
F-statistic:  1141 on 1 and 233 DF,  p-value: < 2.2e-16
```

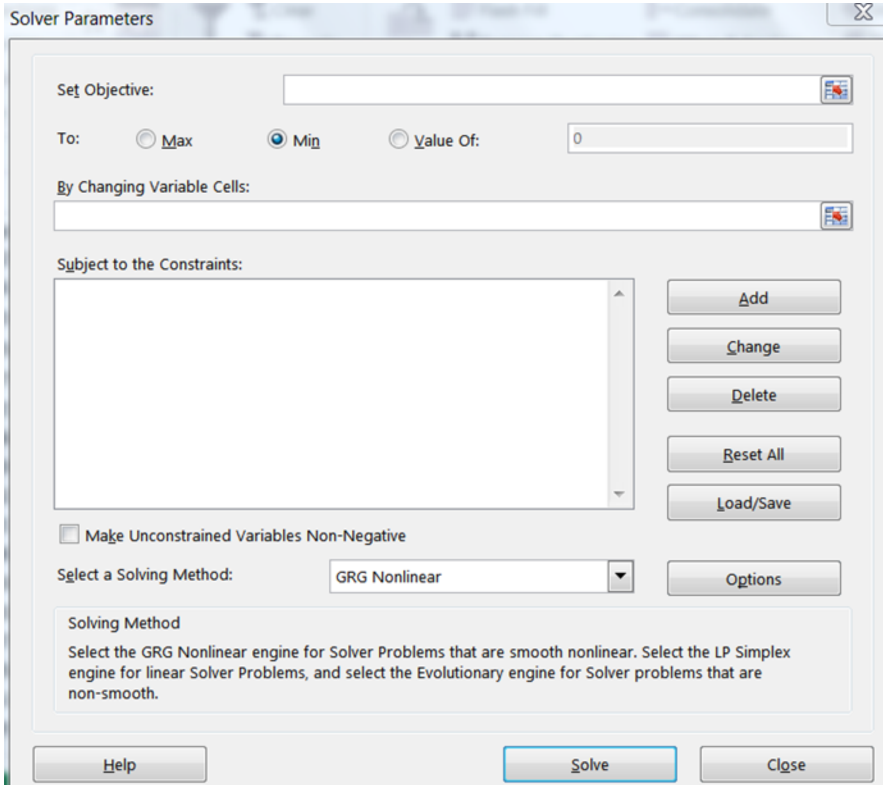
Source: Chew C.H. (2020) AI, Analytics and Data Science Vol. 2.

Class Activity 3

Total Loss = Sum of Check Function

Est. Duration: 20 mins

- Use Excel > Data > Solver to solve for the optimal value of b_0 and b_1 in the Engel Dataset, for various values of Tau τ in Quantile Regression.
  - Define the Total Loss metric for Quantile Regression.
  - Do you get the same answers from Solver compared to R?
  - If different, which answer is better?



Solver Parameters

Set Objective:

To: ○ Max ● Min ○ Value Of: 0

By Changing Variable Cells:

Subject to the Constraints:

□ Make Unconstrained Variables Non-Negative

Select a Solving Method: GRG Nonlinear

Solving Method

Select the GRG Nonlinear engine for Solver Problems that are smooth nonlinear. Select the LP Simplex engine for linear Solver Problems, and select the Evolutionary engine for Solver problems that are non-smooth.

Ans to Class Activity 3

Linear vs Quantile Regression

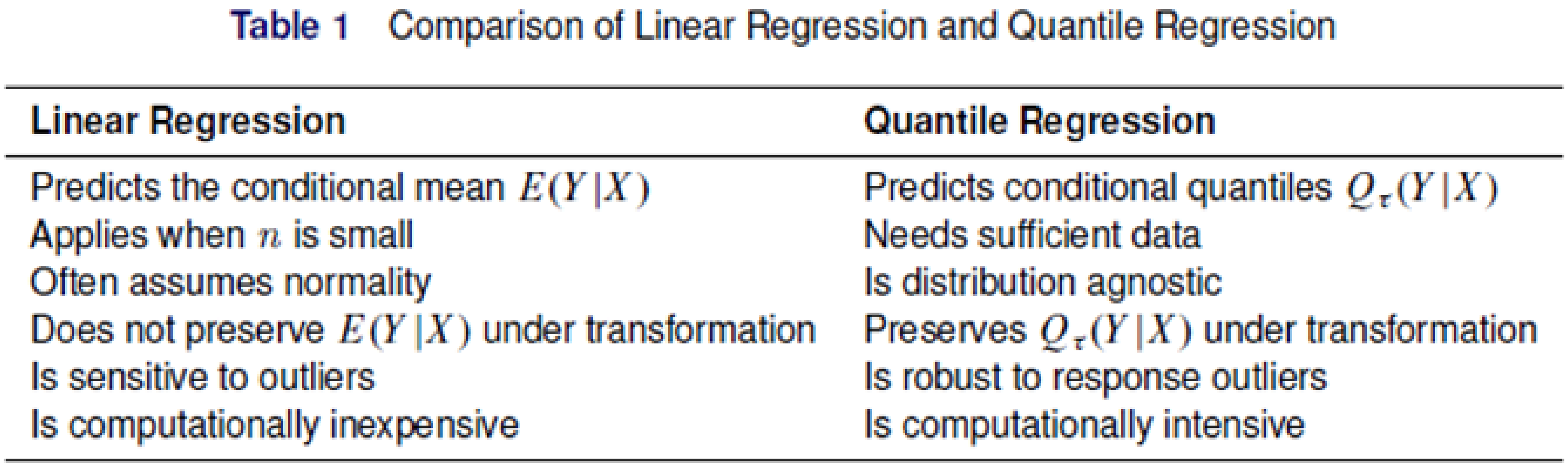


Table 1 Comparison of Linear Regression and Quantile Regression

Linear Regression	Quantile Regression
Predicts the conditional mean $E(Y\|X)$	Predicts conditional quantiles $Q_\tau(Y\|X)$
Applies when n is small	Needs sufficient data
Often assumes normality	Is distribution agnostic
Does not preserve $E(Y\|X)$ under transformation	Preserves $Q_\tau(Y\|X)$ under transformation
Is sensitive to outliers	Is robust to response outliers
Is computationally inexpensive	Is computationally intensive

Source: Rodriguez and Yao (2017) Five Things You Should Know about Quantile Regression. Paper SAS525-2017, SAS Institute.

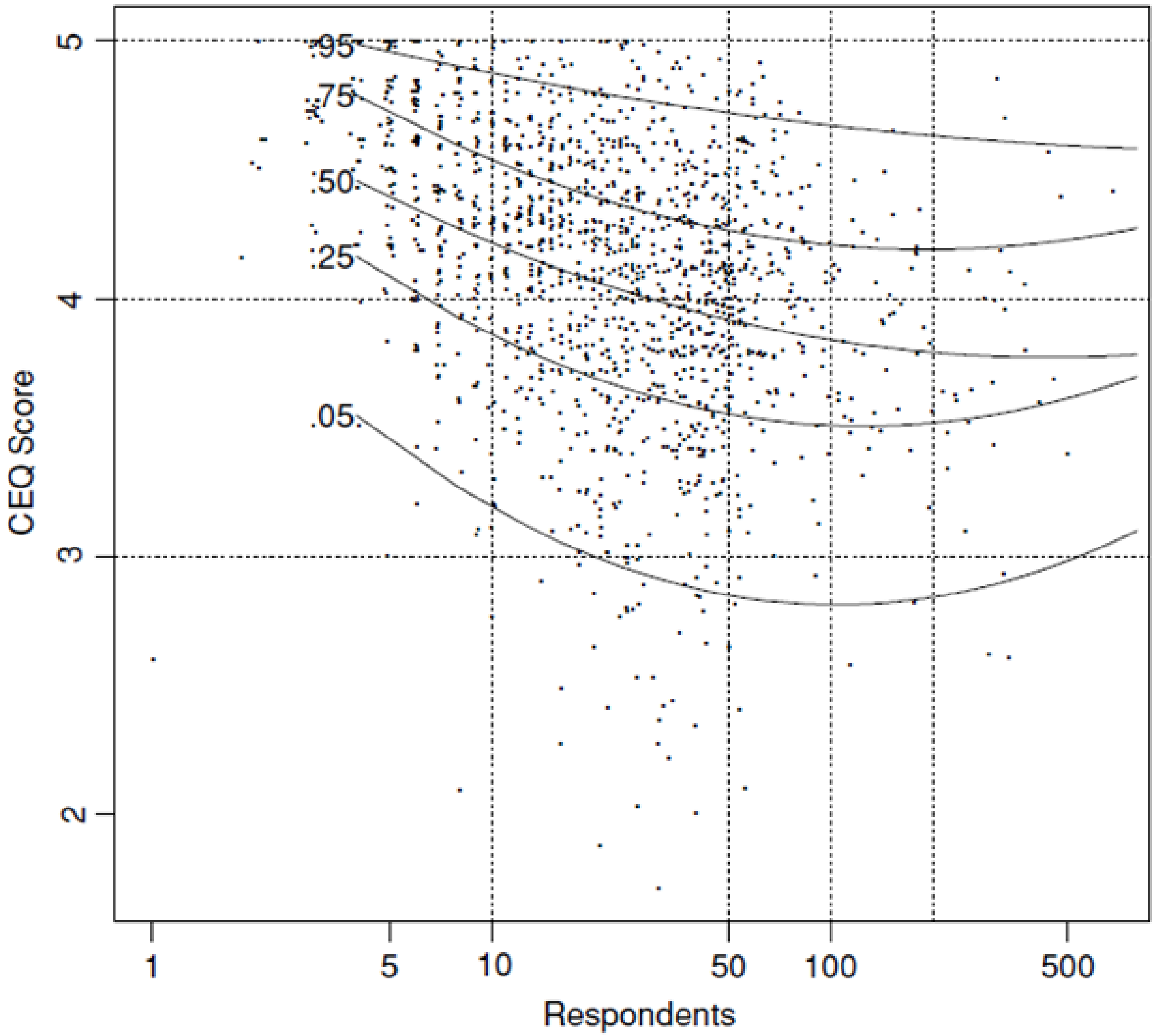
Quantile Regression with Multiple Xs

Quantile Regression with multiple Input Variables

- Example: Effect of class size on course evaluation questionnaire (CEQ) score.
  - The data consist of mean course evaluation scores for 1482 courses offered by a large public university over the period 1980–94.
  - Some courses are undergraduate, some are postgraduate.
  - Class sizes vary.
  - Some classes have good, experienced instructors, some instructors are fresh graduate and has no/limited teaching or working experience.
  - Primarily want to understand the impact of class size on teaching evaluation, despite all the variables.

Source: Roger Koenker (2005). Quantile Regression. Cambridge University Press.

Evaluation Scores with Quantile Regression Lines for a university course with different class size



Source: Roger Koenker (2005). Quantile Regression. Cambridge University Press.

Proposed Quantile Regression Model

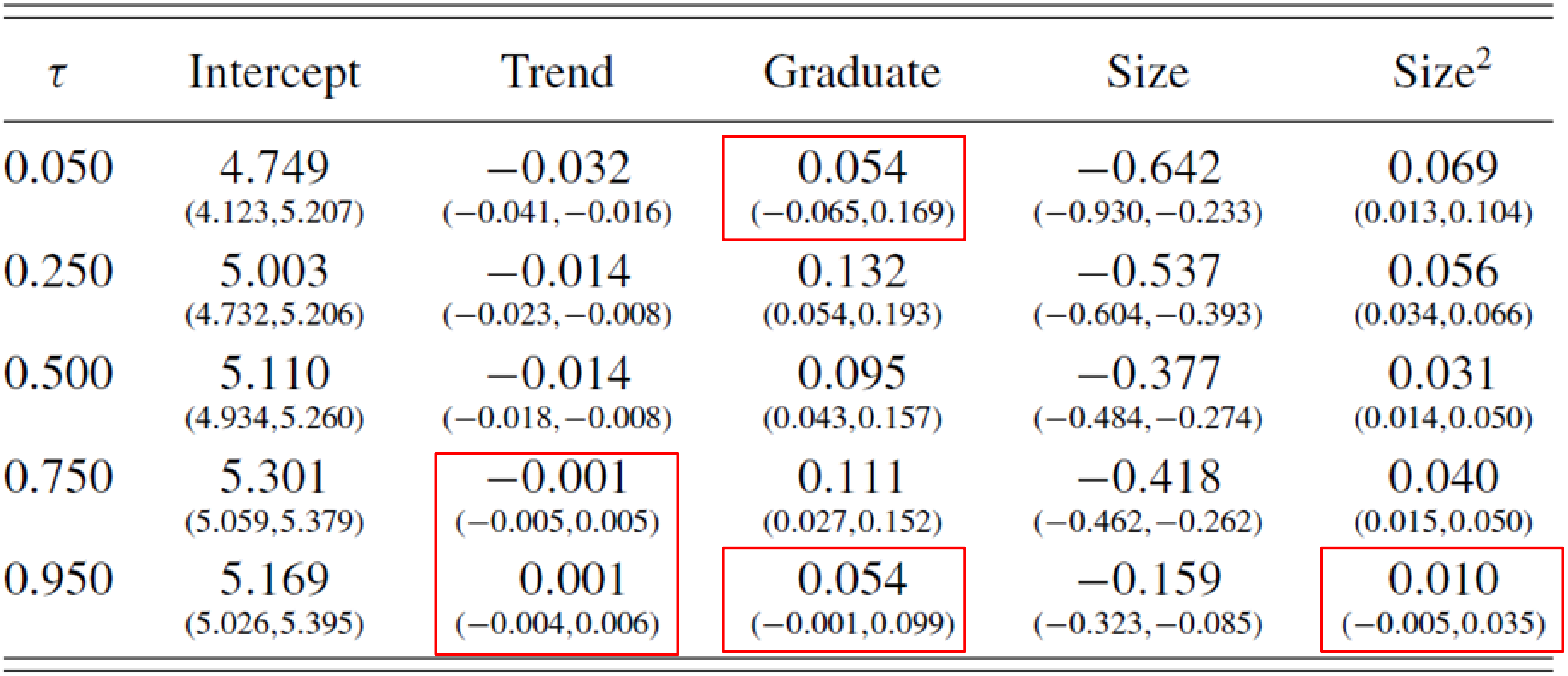
$$Q_Y(\tau|x) = \beta_0(\tau) + \text{Trend}\,\beta_1(\tau) + \text{Grad}\,\beta_2(\tau) + \text{Size}\,\beta_3(\tau) + \text{Size}^2\beta_4(\tau)$$

- Trend: Linear Time Trend Component.
- Grad: 0 (if undergraduate course) or 1 (if postgraduate course).
- Size: number of students in the class.

Source: Roger Koenker (2005). Quantile Regression. Cambridge University Press.

Quantile Regression Model Results

$$Q_Y(\tau|x) = \beta_0(\tau) + \text{Trend}\,\beta_1(\tau) + \text{Grad}\,\beta_2(\tau) + \text{Size}\,\beta_3(\tau) + \text{Size}^2\beta_4(\tau)$$



τ	Intercept	Trend	Graduate	Size	Size^2
0.050	4.749 (4.123,5.207)	−0.032 (−0.041,−0.016)	0.054 (−0.065,0.169)	−0.642 (−0.930,−0.233)	0.069 (0.013,0.104)
0.250	5.003 (4.732,5.206)	−0.014 (−0.023,−0.008)	0.132 (0.054,0.193)	−0.537 (−0.604,−0.393)	0.056 (0.034,0.066)
0.500	5.110 (4.934,5.260)	−0.014 (−0.018,−0.008)	0.095 (0.043,0.157)	−0.377 (−0.484,−0.274)	0.031 (0.014,0.050)
0.750	5.301 (5.059,5.379)	−0.001 (−0.005,0.005)	0.111 (0.027,0.152)	−0.418 (−0.462,−0.262)	0.040 (0.015,0.050)
0.950	5.169 (5.026,5.395)	0.001 (−0.004,0.006)	0.054 (−0.001,0.099)	−0.159 (−0.323,−0.085)	0.010 (−0.005,0.035)

Source: Roger Koenker (2005). Quantile Regression. Cambridge University Press.

Quantile Regression Model Conclusion (from the Results table)

- What is common among the red boxes in the teaching evaluation score quantile regression results? What does this mean in the business context?

- Provide 3 conclusions from the results table.

Quantile Regression Model Conclusions

Summary

- Quantile Regression:
  - Specify Percentile of Interest.
  - Allows more flexibility in modelling data, compared to linear regression.
  - Rpackage: quantreg
  - Python: statsmodels
  - SAS Stat Procedure: quantreg

Reminder

Please complete the Pre-Class Learning Activities before next class.

Reflection on your Learning

Go

NTULearn Class Site > Journal

Post

Read the instructions and post entry on this week’s learning.

- Reply on the 3 questions as stated in the Journal Instructions.