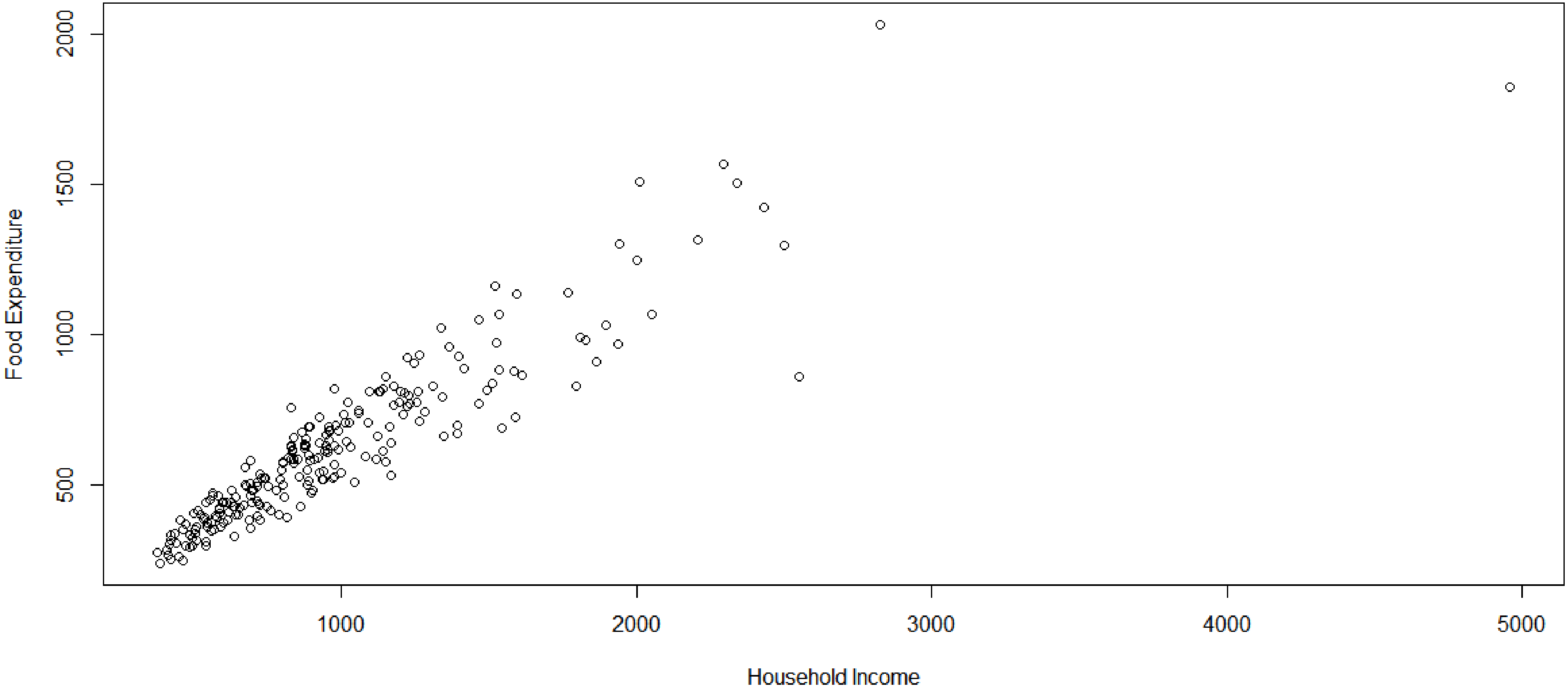
Answers in Quantile Regression

BC2407 ANALYTICS II SEMINAR 4

NEUMANN CHEW C. H.

What is the (business) purpose of analyzing this data?



To study how family expenditure on food is affected by income (i.e. cost of living).

Review of Linear Regression Model Line

$$y = \underbrace{b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_m x_m}_{\hat{y}} + e$$

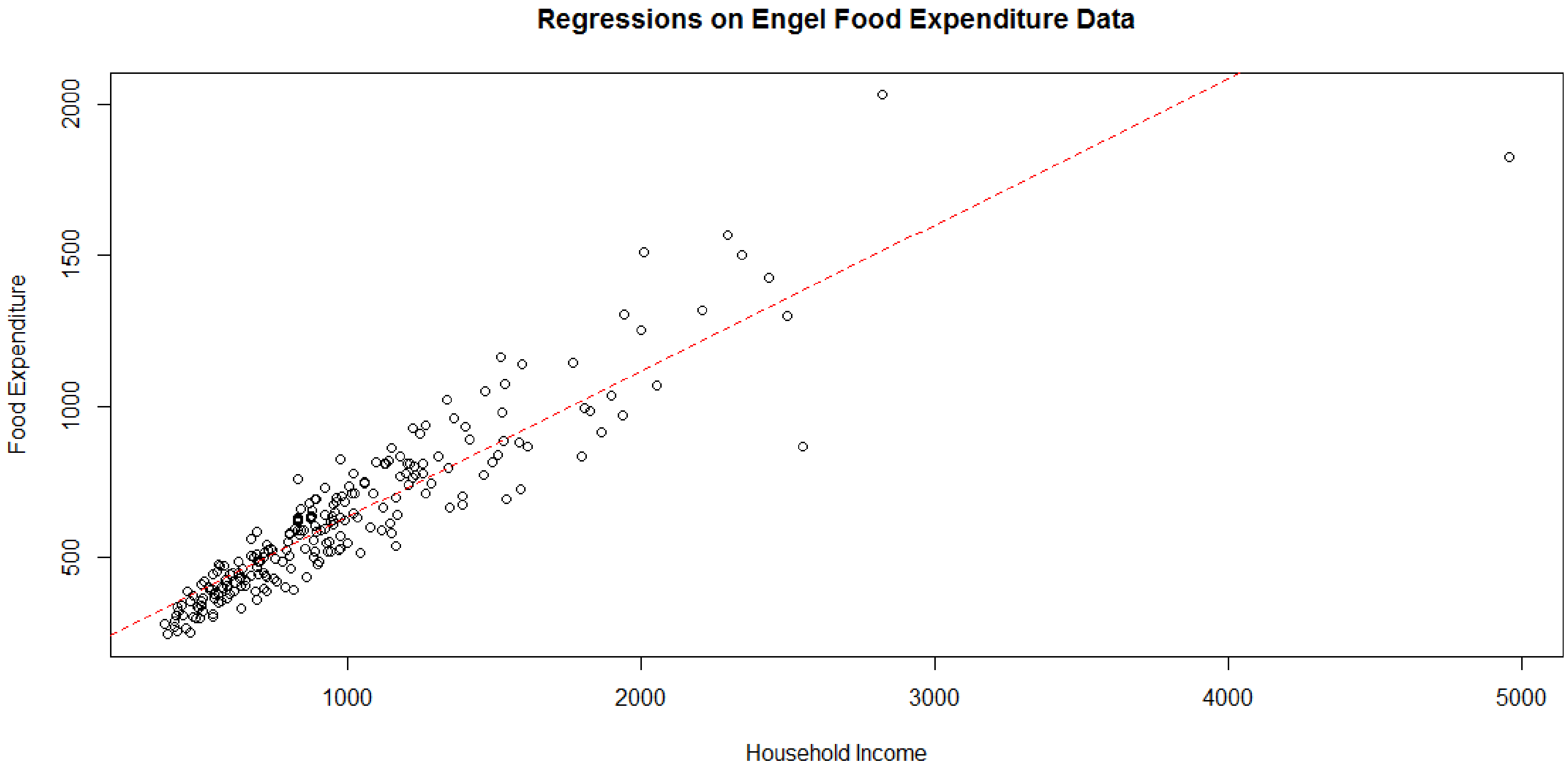
Straight Line Equation

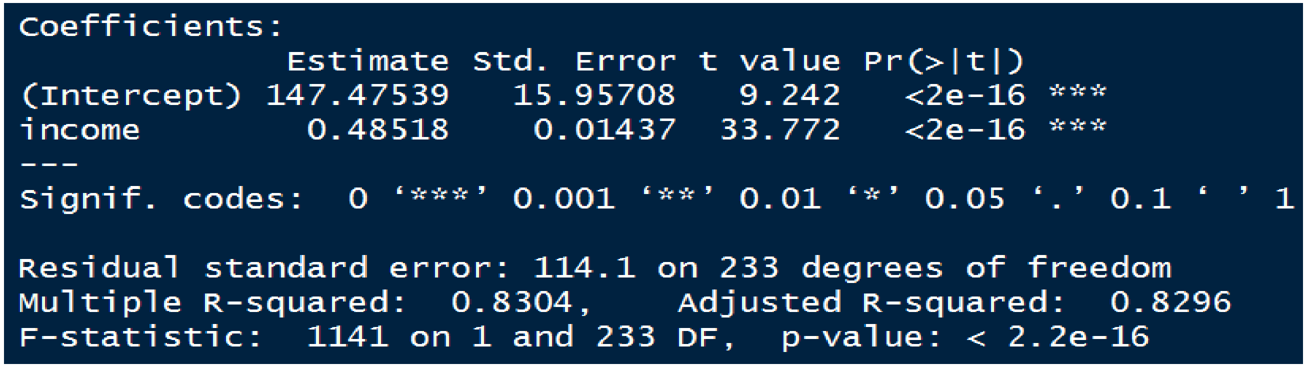
$e \sim N(0, \sigma)$

Errors (aka Residuals) follow a Normal Distribution with mean 0 and constant standard deviation.

Q: What does the straight line equation actually represent?

A: The mean value of Y, at the specified value of Xs.



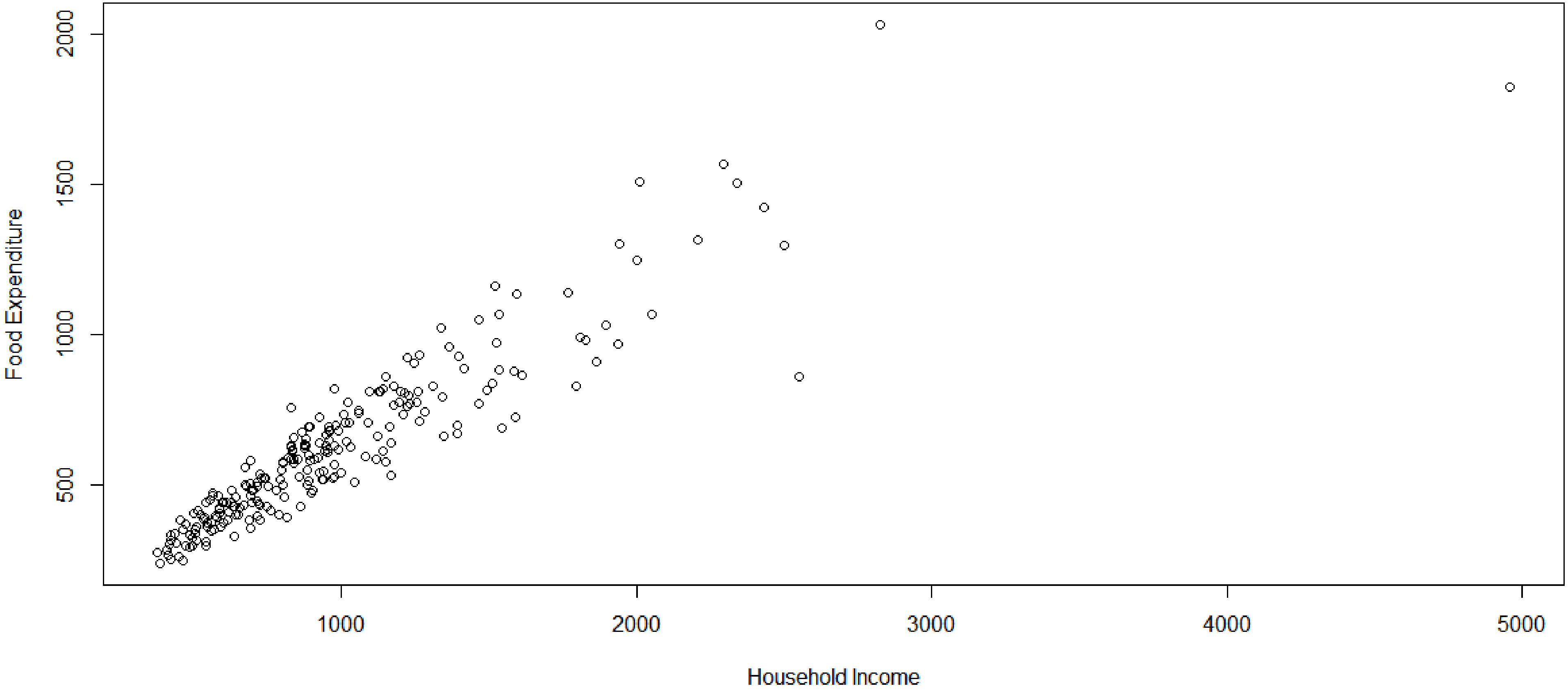


```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 147.47539   15.95708   9.242   <2e-16 ***
income        0.48518    0.01437  33.772   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

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
```

P-value is very low => Income is a significant predictor of Food Expenditure

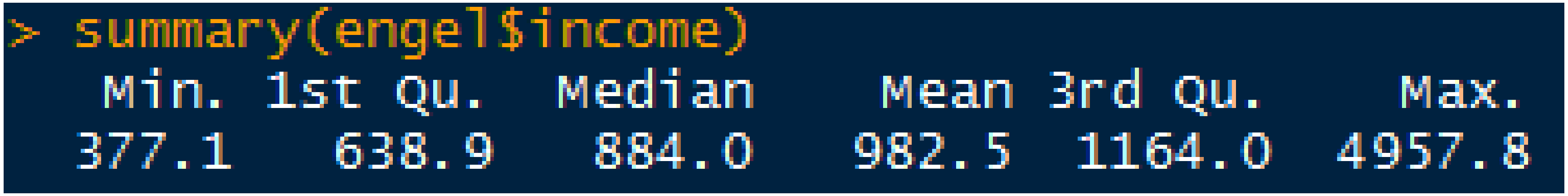
Specific Business Questions to be Answered?



How much does a typical family spend on Food?

More specific Business Questions

- How much does a typical family spend on Food?
  - What do you mean by typical family?
    - Family with mean income?
    - Is this the only kind of family that one is interested in analysing?



```
> summary(engel$income)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  377.1   638.9   884.0   982.5  1164.0  4957.8 
```

- “Typical” Low Income family. e.g. $500
- “Typical” Average Income family. e.g. $1000
- “Typical” High Income family. e.g. $2500

Class Activity 1

Quantile Regression in R

Est. Duration: 20 mins

3:16pm – 3:36pm

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?

Ans: False. Check degrees of freedom in the software output.

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

Class Activity 2

Total Loss in 3 Models for Each Tau

Est. Duration: 20 mins

4:35pm – 4:55pm

- 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

1. See result in Excel File Check Function solutions.xlsx
2. Yes, all data points are used, regardless of tau value.
3. The higher the tau value, the higher the height of the quantile regression line as the total loss will be lower.

Tau at middle values require more data points to disambiguate [clear winning model] compared to Tau at extreme values (e.g. 0.1 or 0.9).

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

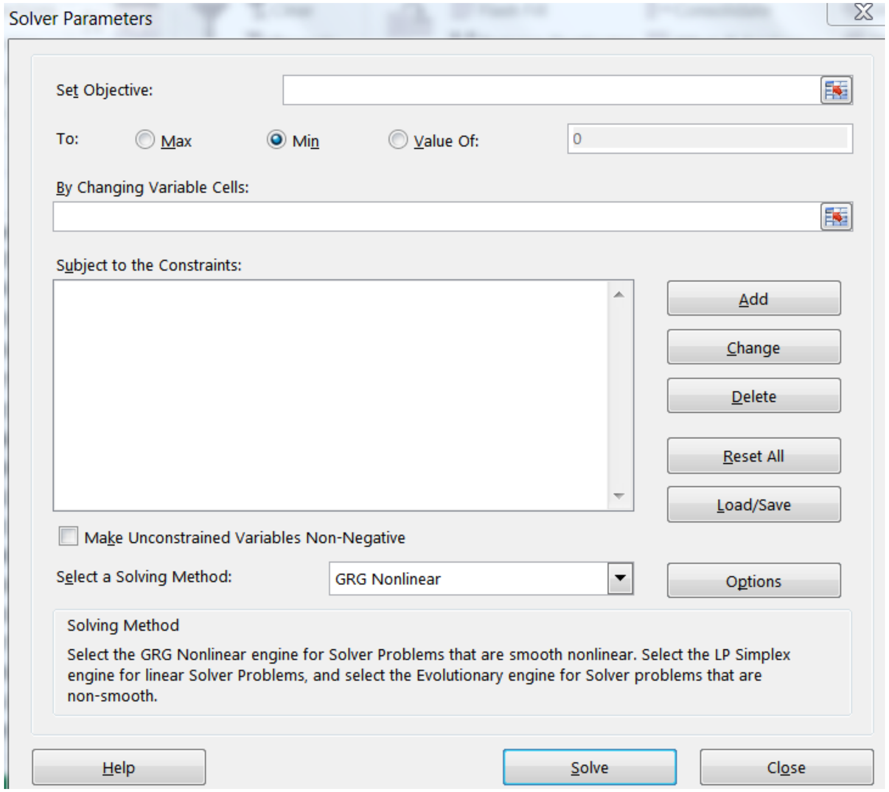
Class Activity 3

Total Loss = Sum of Check Function

Est. Duration: 20 mins

5:01pm – 5:21pm

- 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

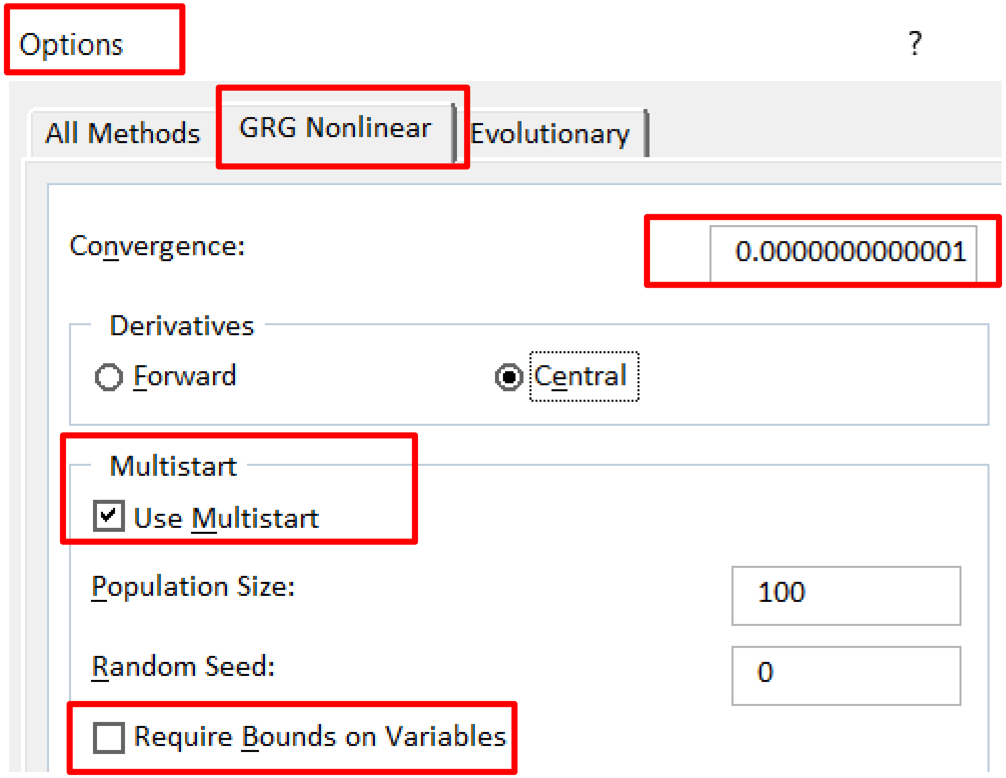
Select the GRG Nonlinear engine for Solver Problems that are smooth nonlinear. Select the LP Simplex engine for linear Solver Problems, and select the Evolutionary engine for Solver problems that are non-smooth.

Ans to Class Activity 3

- Given the equations and solver settings, as-is, the Solver results are different compared to R.
  - Adjust the Solver Options to get closer to global optimal solution

- The parameter results from R are better as they produce a smaller total loss compared to Solver parameter values.

- The learning objective is to understand how Quantile Regression Model is formed by minimizing the Total Loss = Sum of the Check Function value for each data point.



Options

Convergence: 0.0000000000001

Derivatives

○ Forward ● Central

Multistart

☑ Use Multistart

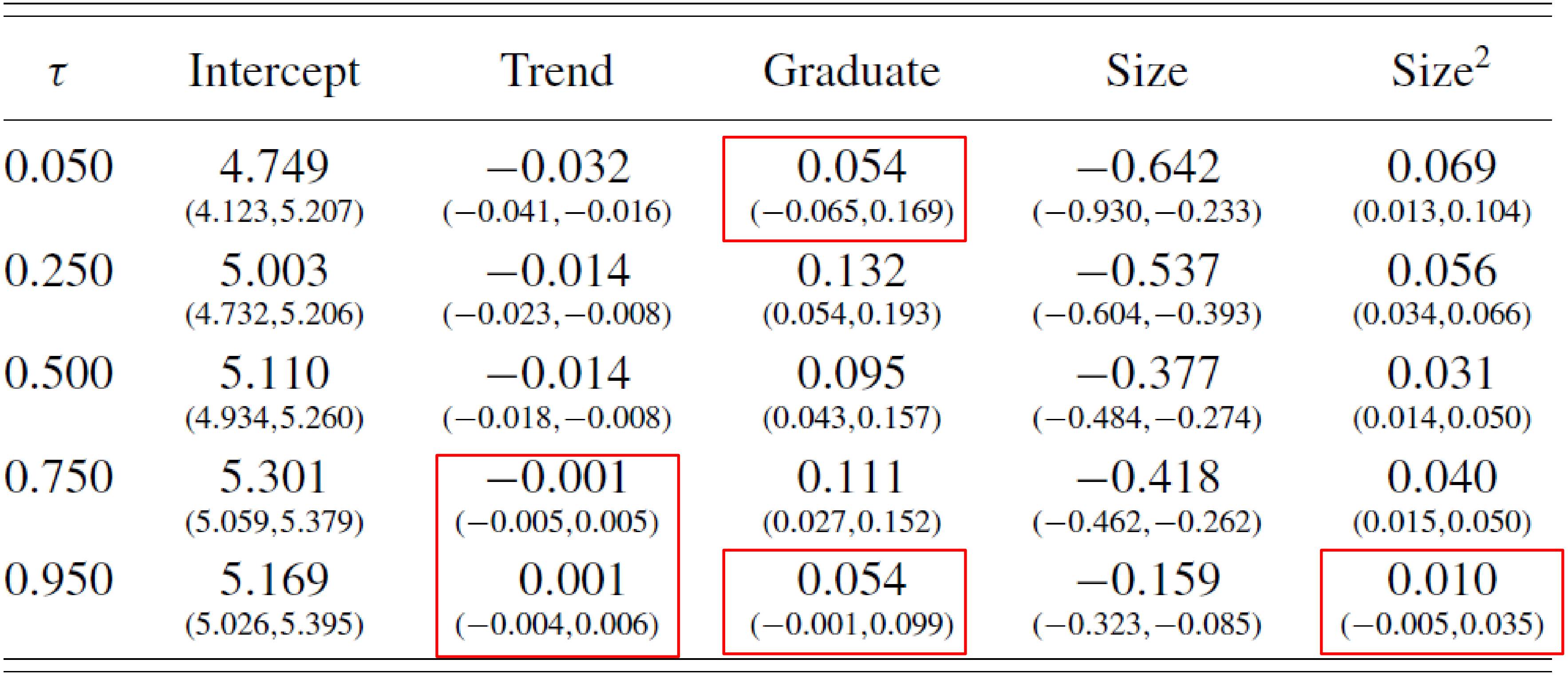
Population Size: 100

Random Seed: 0

□ Require Bounds on Variables

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)

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?

  Ans: Confidence Interval includes 0! This implies that those factors are not statistically significant in affecting teaching evaluation scores, at that teaching score percentile.

- Provide 3 conclusions from the results table.

Quantile Regression Model Conclusions

1. The downward trend in teaching ratings is statistically significant at the median and lower quantiles i.e. average and poor ratings are getting worse.
2. Among the best and worst teaching evaluations, it does not matter whether the class is undergraduate or graduate. The type of class is only statistically significant for the median and near to median teaching ratings. The ratings would be higher for graduate classes.
3. Larger classes has lower ratings regardless of teaching ratings. However, all except the best ratings have a turning point at certain big enough class size.
   - School deployed experienced professors to teach very large classes.