NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Quantile Regression

BC2407 ANALYTICS II SEMINAR 4

NEUMANN CHEW C. H.

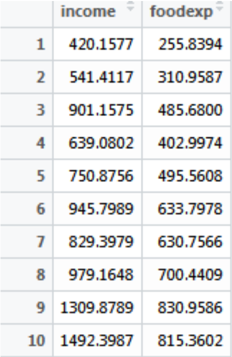
From Linear Regression…

- Recall assumptions of Linear Regression.
  1. Y has a linear relationship with the Xs.
  2. Residuals has a Normal distribution.
  3. Residuals has constant variance independent of Xs.
- Checked via 4 Diagnostic Plots in R with plot().
- What if one or more assumptions are not satisfied?
  - Knowing only linear regression, we can only try to do mathematical transformations of the variables and try linear reg on transformed variables e.g. sqrt(X_1), log(Y).
  - Quantile Reg provide a natural alternative to Linear Reg if assumption 3 is violated.

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

Engel Dataset from Rpackage quantreg

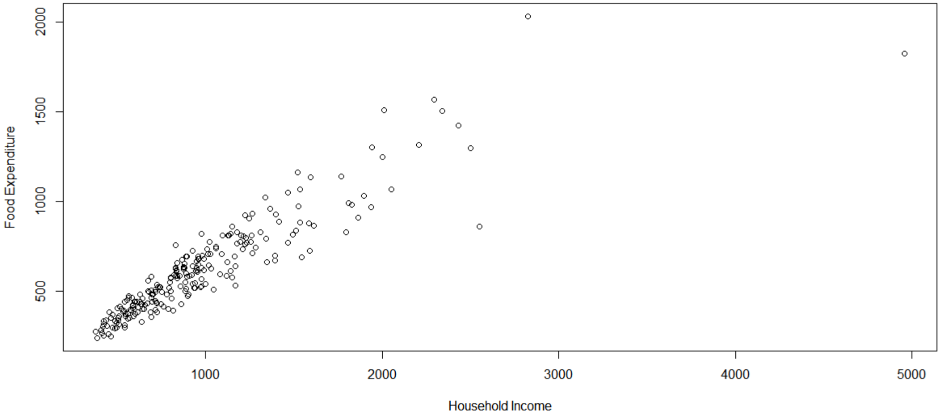
- Dataset that records Family Expenditure on Food and Family Income in Belgium 1857.
- Used to show a limitation of Linear Regression and usefulness of Quantile Regression.
- Dataset is in quantreg Rpackage.



	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.

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

$$y = \underbrace{b_0 + b_1x_1 + b_2x_2 + \cdots + b_mx_m}_{\hat{y}} + e$$

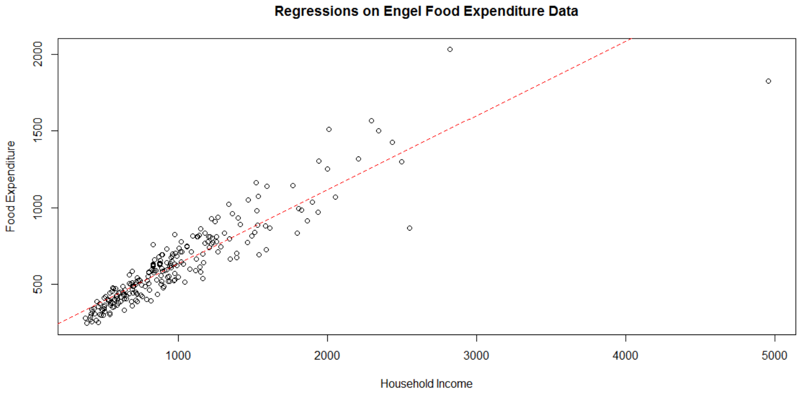
e ~ N(0, σ)

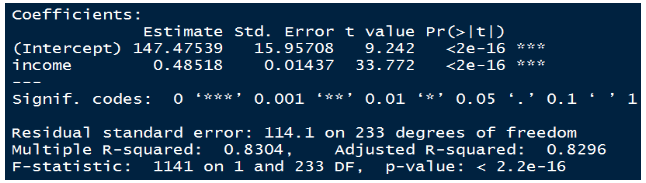
Straight Line Equation

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.

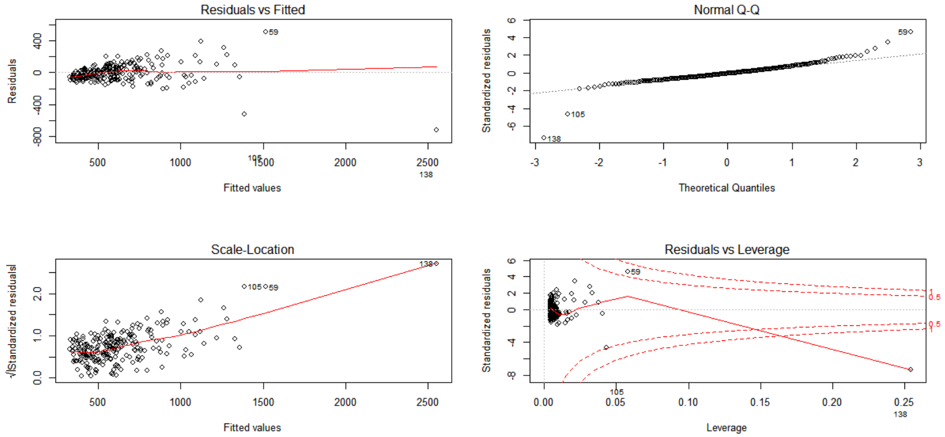




```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) 147.47539   15.95708   9.242   <2e-16 ***
income        0.48518    0.01437  33.772   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

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

Diagnostic Plots reveal problems



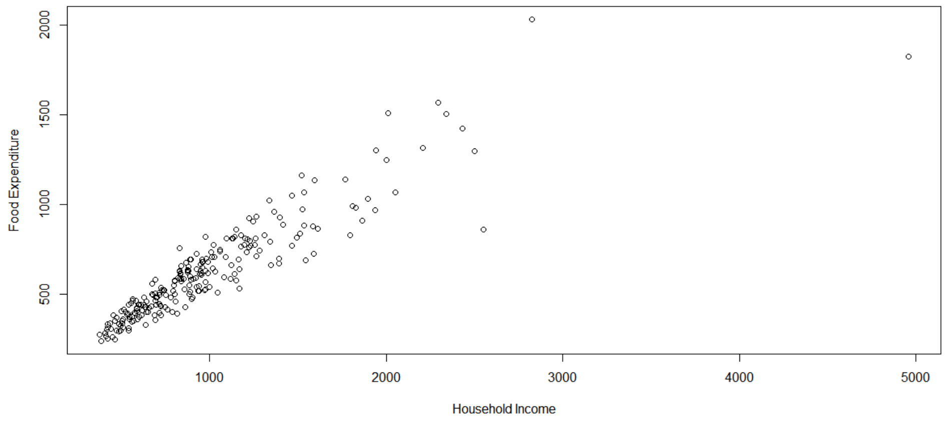
Error increases at higher predicted expenditures.

Influential outlier detected

Conclusion of Linear Regression Diagnostics

- Linear Regression Model Assumptions are not met.
- Still proceed with Linear Regression?
- Let’s relook the data. What do you want to find out? (in the business/social sense, not mathematics.)

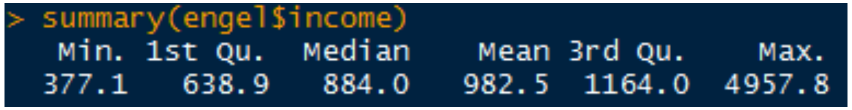
Specific Business Questions to be Answered?



How much does a typical family spend on Food?

More specific Business Questions

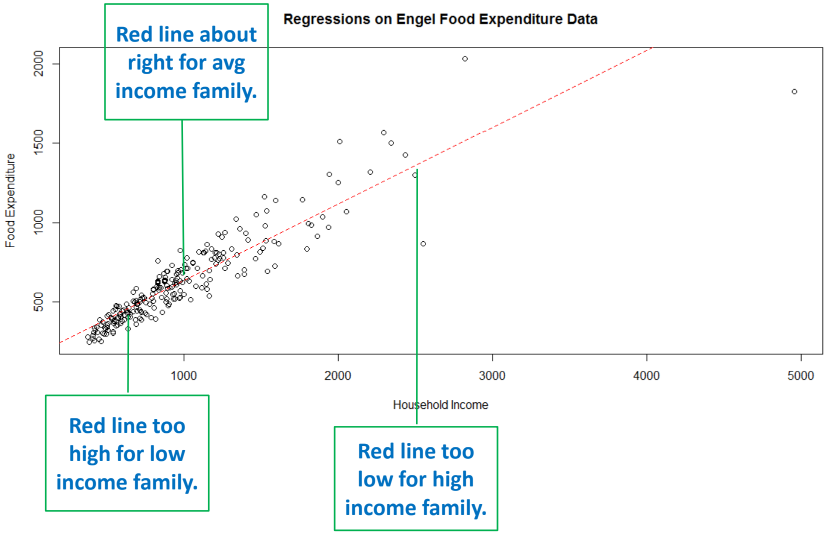
- How much does a typical family spend on Food?
  - What do you mean by typical family?
    - Family with mean income?
    - Is this the only kind of family that one is interested in analysing?



```
> 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

Linear Regression (Dotted Red Line) is Inadequate for answering questions about “typical” family expenditure



Red line about right for avg income family.

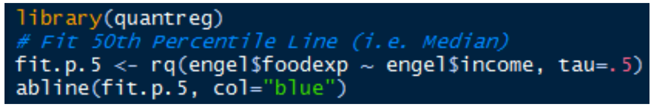
Red line too high for low income family.

Red line too low for high income family.

Other Cases where Linear Regression is Inadequate

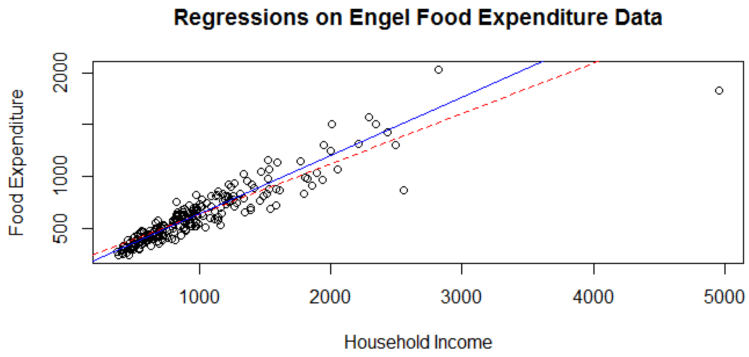
- If we aim to hit higher for target variable Y (e.g. productivity, profits,…), then need to find out and aim for the 95th or 99th percentile of Y.
- If we aim to hit lower for target variable Y (e.g. waiting time, losses,…), then need to find out and aim for the 5th or 1st percentile of Y.
- If Y is highly skewed, then a better measure of the “average” value of Y is the median of Y (50th percentile), instead of mean.
- i.e. Knowing the mean of Y [e.g. the dotted red line] is often inadequate in such cases.

rq() function in quantreq Rpackage. Median represented as Blue Line Mean as dotted red line.

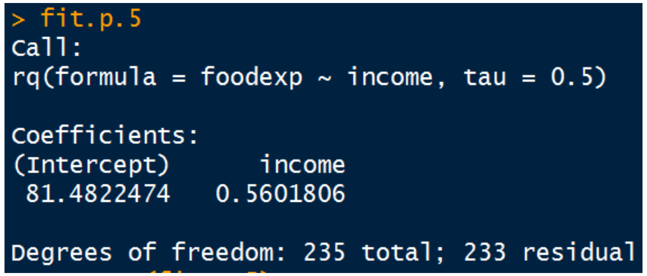


```
library(quantreg)
# Fit 50th Percentile Line (i.e. Median)
fit.p.5 <- rq(engel$foodexp ~ engel$income, tau=.5)
abline(fit.p.5, col="blue")
```

Parameter tau controls the percentile.

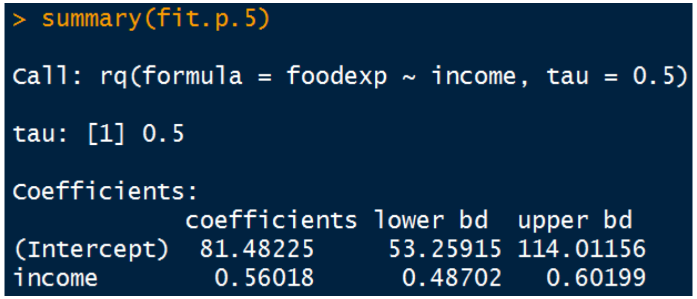


Run the name of the model to get model coefficients of the Quantile Regression Model



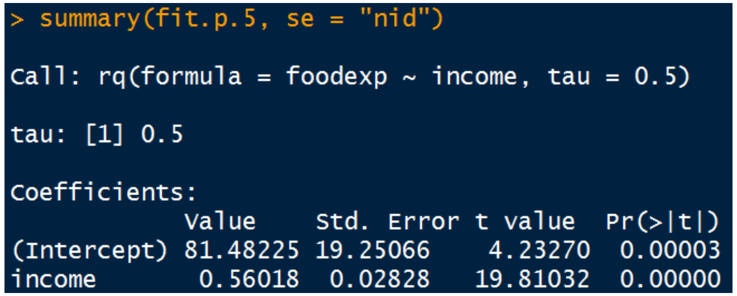
```
> fit.p.5
Call:
rq(formula = foodexp ~ income, tau = 0.5)

Coefficients:
(Intercept)      income
 81.4822474   0.5601806

Degrees of freedom: 235 total; 233 residual
```

Use summary() to get model coefficients and confidence intervals of the Quantile Regression model



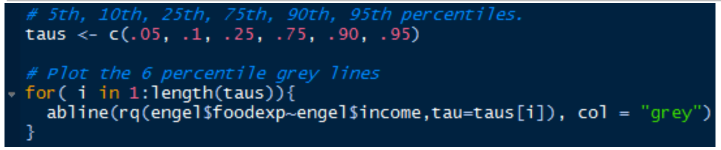
```
> summary(fit.p.5)

Call: rq(formula = foodexp ~ income, tau = 0.5)

tau: [1] 0.5

Coefficients:
            coefficients lower bd  upper bd 
(Intercept) 81.48225     53.25915 114.01156
income       0.56018      0.48702   0.60199
```

Use summary(..., se = “nid”) to get P-values of the Quantile Regression model coefficients

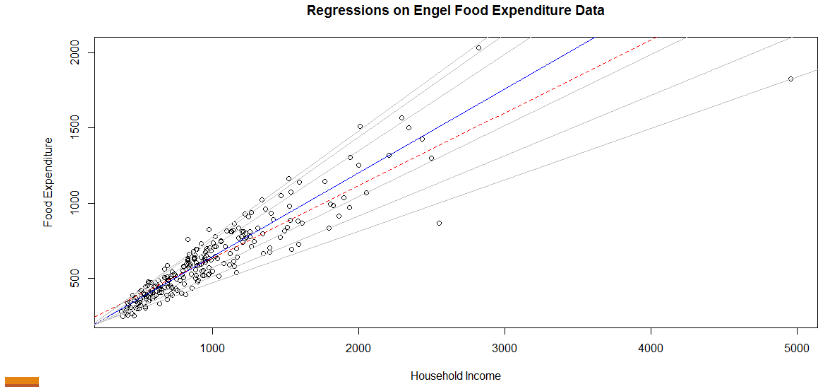


```
> summary(fit.p.5, se = "nid")

Call: rq(formula = foodexp ~ income, tau = 0.5)

tau: [1] 0.5

Coefficients:
            Value    Std. Error t value  Pr(>|t|)
(Intercept) 81.48225 19.25066    4.23270  0.00003
income       0.56018  0.02828   19.81032  0.00000
```

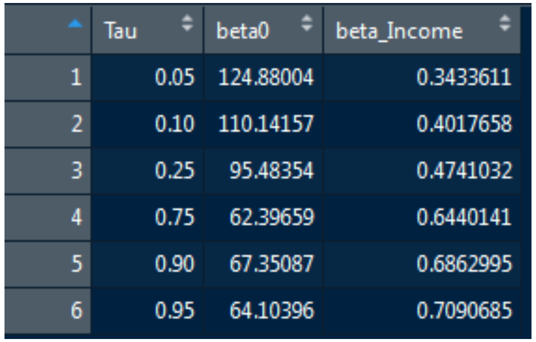
Fit & Plot Six Other Percentiles Quickly with a for loop



```
# 5th, 10th, 25th, 75th, 90th, 95th percentiles.
taus <- c(.05, .1, .25, .75, .90, .95)

# Plot the 6 percentile grey lines
for( i in 1:length(taus)){
  abline(rq(engel$foodexp~engel$income,tau=taus[i]), col = "grey")
}
```



Quantile Regression Model Coefficients at various Percentiles



	Tau	beta0	beta_Income
1	0.05	124.88004	0.3433611
2	0.10	110.14157	0.4017658
3	0.25	95.48354	0.4741032
4	0.75	62.39659	0.6440141
5	0.90	67.35087	0.6862995
6	0.95	64.10396	0.7090685

Class Activity 1

Quantile Regression in R

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?

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

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

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.

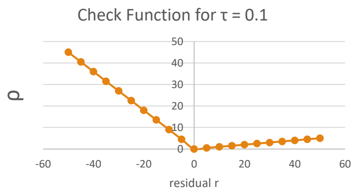
Understanding the Check Function

USED IN QUANTILE REGRESSION

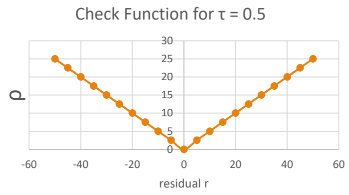
Understanding the Check Function:

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

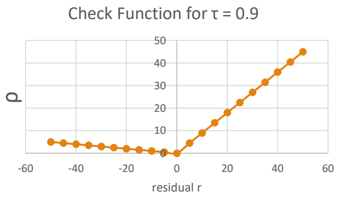
- See Excel File: Check Function > By Tau, for examples using 3 values of τ.
- What can you conclude from the numerical examples?



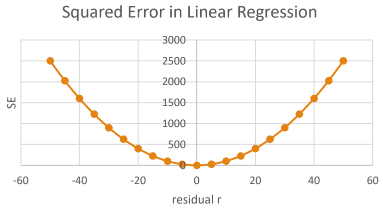
Check Function for τ = 0.1



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.

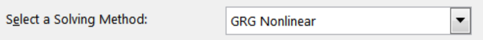
Ans to Class Activity 2

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.
4. 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.

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

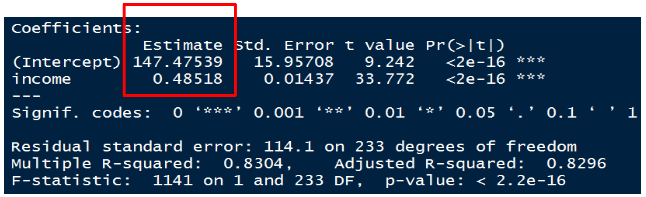
- Plan:
  - Define the Total Loss metric to be minimized.
  - Use Excel Solver to find the values for the linear regression model coefficients b_0 and b_1 that will minimize the Total Loss.
  - Select solving method to GRG Nonlinear



Select a Solving Method: GRG Nonlinear

Note: If Solver is missing in Excel, activate 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
```

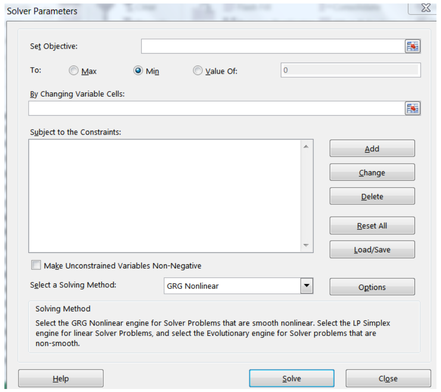
Source: Chew C.H. (2020) AI, Analytics and Data Science Vol. 2.

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

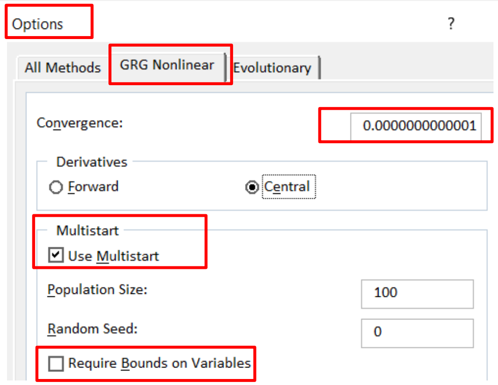
Select a Solving Method: GRG Nonlinear

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

- 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

All Methods | GRG Nonlinear | Evolutionary

Convergence: 0.0000000000001

Derivatives
○ Forward ● Central

Multistart
☑ Use Multistart

Population Size: 100

Random Seed: 0

□ Require Bounds on Variables

Linear vs Quantile Regression

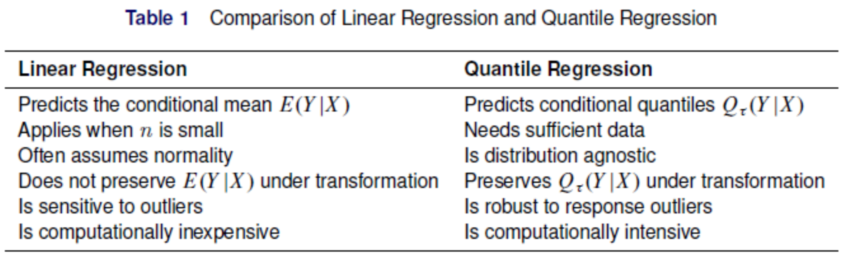


Table 1 Comparison of Linear Regression and Quantile Regression

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 using Python

Quantile Regression in Python

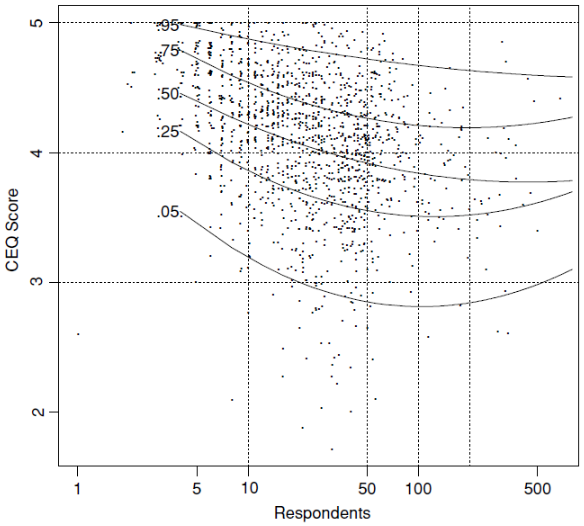
- statsmodels
  - Includes Engel dataset
  - https://www.statsmodels.org/dev/examples/notebooks/generated/quantile_regression.html
- There are other packages that claim to be able to execute quantile regression (e.g. mlinsights).

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.

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

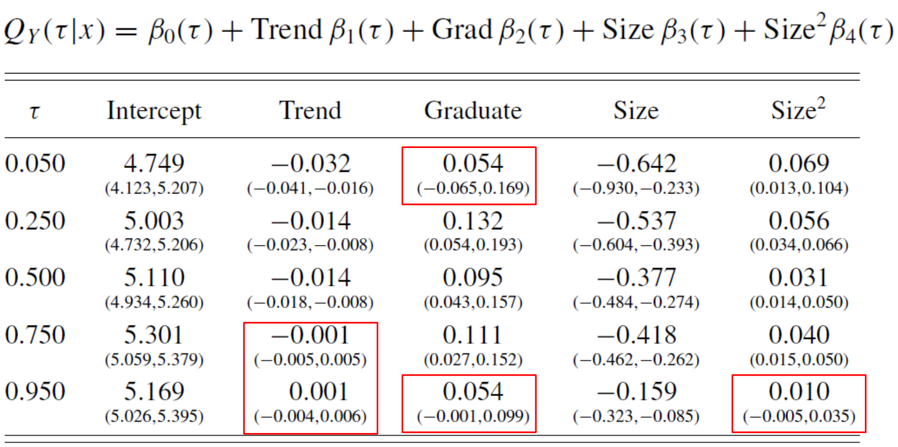


Proposed Quantile Regression Model

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

- Trend: Last year teaching evaluation score.
- Grad: 0 (if undergraduate course) or 1 (if postgraduate course).
- Size: number of students in the class.

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.

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

Check Your Understanding

- Review Questions