Quantile Regression

Justin Raymond S. Eloriaga

2021

Chapter Summary

We show the effectivity and flexibility of quantile regressions as compared to traditional linear models. We derive the LAD estimate and generalize this for all quantiles. We then show this graphically and highlight the key advantages of the quantile regression.

What's under attack now?

What's under attack now?

- I mentioned in past meetings that the reason why other models exists is simply because CLRM isn't perfect. So what assumptions are we targeting now?

What's under attack now?

- I mentioned in past meetings that the reason why other models exists is simply because CLRM isn't perfect. So what assumptions are we targeting now?
- Well one assumption we made is that errors are normally distributed. And that may be hard to meet in real life.

What's under attack now?

- I mentioned in past meetings that the reason why other models exists is simply because CLRM isn't perfect. So what assumptions are we targeting now?
- Well one assumption we made is that errors are normally distributed. And that may be hard to meet in real life.
- We also showed how outliers can throw off our estimation, and in turn, may cause us to make misleading inferences.

Food and Income

Food and Income

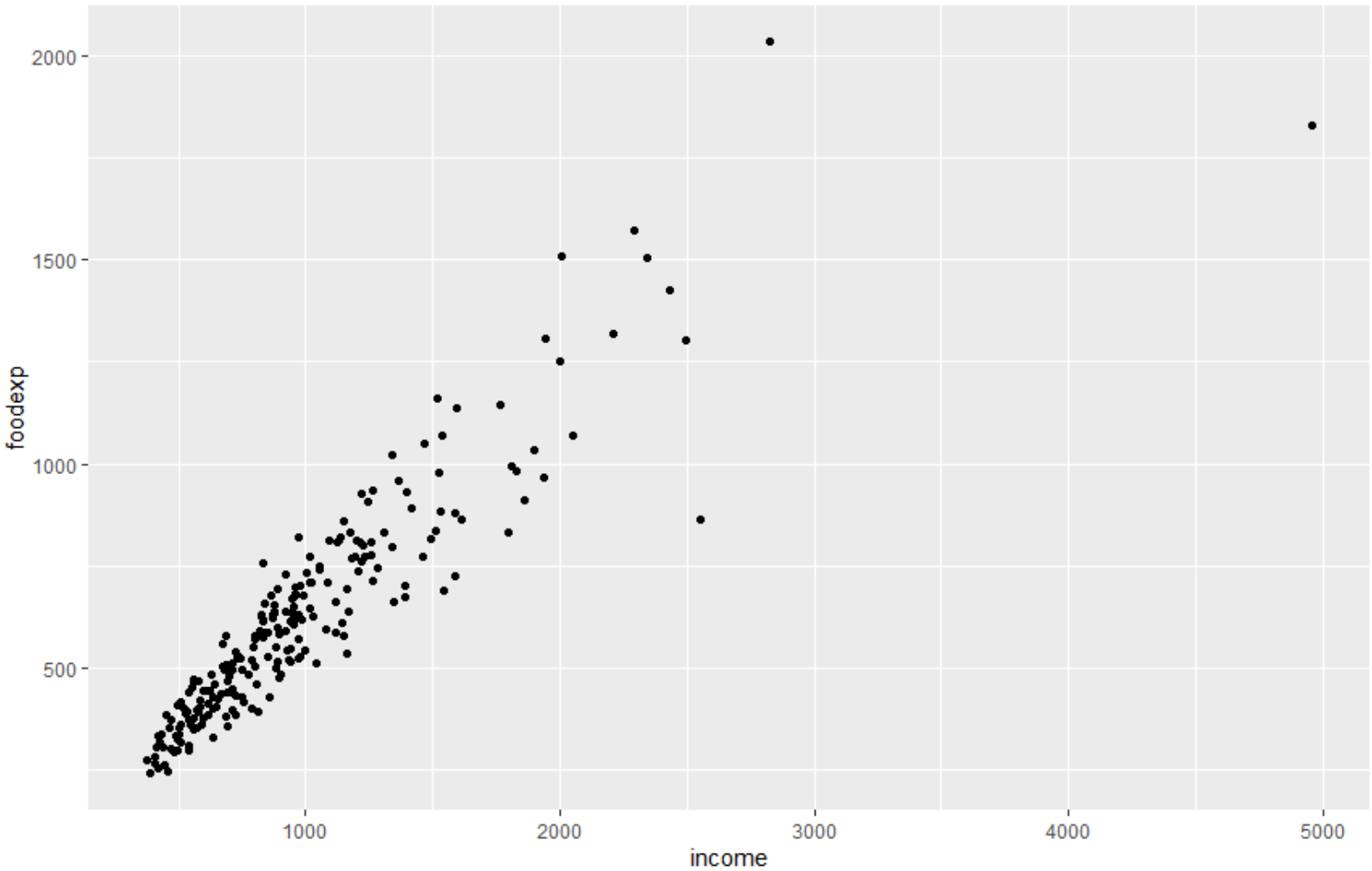
Most millenials/Gen Z's can attest to this meme...

Food and Income

Most millenials/Gen Z's can attest to this meme...

Let's try and zero in on the relationship between food and income.

Let's Start with a Scatterplot



Initial Thoughts

Initial Thoughts

If we run a model like $FoodExp_i = \beta_0 + \beta_1 Income_i + u_i$, we would likely conclude that $\beta_1 > 0$. We obviously see a positive relationship between food expenditure and income.

Initial Thoughts

If we run a model like $FoodExp_i = \beta_0 + \beta_1 Income_i + u_i$, we would likely conclude that $\beta_1 > 0$. We obviously see a positive relationship between food expenditure and income.

- But what if we wanted to see it for those in the population that have higher incomes? Is that relationship stronger or weaker (i.e. is β_1 more positive or less positive?

Initial Thoughts

If we run a model like $FoodExp_i = \beta_0 + \beta_1 Income_i + u_i$, we would likely conclude that $\beta_1 > 0$. We obviously see a positive relationship between food expenditure and income.

- But what if we wanted to see it for those in the population that have higher incomes? Is that relationship stronger or weaker (i.e. is β_1 more positive or less positive?
- Traditionally, we have relied on interaction terms (i.e. the product of a dummy and a continous variable to alter the slope.

Initial Thoughts

If we run a model like $FoodExp_i = \beta_0 + \beta_1 Income_i + u_i$, we would likely conclude that $\beta_1 > 0$. We obviously see a positive relationship between food expenditure and income.

- But what if we wanted to see it for those in the population that have higher incomes? Is that relationship stronger or weaker (i.e. is β_1 more positive or less positive?
- Traditionally, we have relied on interaction terms (i.e. the product of a dummy and a continous variable to alter the slope.
- But these interaction terms may be misleading if some assumptions of the CLRM are not met.

OLS Results

```
Call:
lm(formula = engel$foodexp ~ engel$income)

Residuals:
    Min      1Q  Median      3Q     Max
-725.70  -60.24   -4.32   53.41  515.77

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 147.47539   15.95708   9.242   <2e-16 ***
engel$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
```

Thinking Beyond the Mean

Thinking Beyond the Mean

Your standard CLRM through the use of OLS explains the *average relationship* (mean) between a set of regressors and the dependent variable variable. $E(Y|X)$.

Thinking Beyond the Mean

Your standard CLRM through the use of OLS explains the *average relationship* (mean) between a set of regressors and the dependent variable variable. $E(Y|X)$.

- From our discussions before, the OLS methodology is predicated on the minimization of the residual sum of squares

$$\min RSS = \sum_{i=1}^{N} e_i^2$$

Thinking Beyond the Mean

Your standard CLRM through the use of OLS explains the *average relationship* (mean) between a set of regressors and the dependent variable variable. $E(Y|X)$.

- From our discussions before, the OLS methodology is predicated on the minimization of the residual sum of squares

$$\min RSS = \sum_{i=1}^{N} e_i^2$$

- But we alluded to the question in the discussion? Is there a more straightforward way for us to prevent the problem of offset?

Thinking Beyond the Mean

Your standard CLRM through the use of OLS explains the *average relationship* (mean) between a set of regressors and the dependent variable variable. $E(Y|X)$.

- From our discussions before, the OLS methodology is predicated on the minimization of the residual sum of squares

$$\min RSS = \sum_{i=1}^{N} e_i^2$$

- But we alluded to the question in the discussion? Is there a more straightforward way for us to prevent the problem of offset?
- The answer is yes and we call it the quantile regression!

What is a Quantile?

What is a Quantile?

- A quantile is basically a percentile, they are synonymous.

What is a Quantile?

- A quantile is basically a percentile, they are synonymous.
- In essence, we have 99 quantiles out there, the best known being the 50th quantile or what we refer to as the *median*.

What is a Quantile?

- A quantile is basically a percentile, they are synonymous.
- In essence, we have 99 quantiles out there, the best known being the 50th quantile or what we refer to as the *median*.
- We will allude to the fact that OLS is just a picture of the conditional mean $E(Y|X)$, which is deficient in many ways. The quantile regression allows us to come up with a more comprehensive picture of the effect of the regressors on the dependent variable.

Least Absolute Deviations

Least Absolute Deviations

That more straightforward manner would just be minimizing the sum of the absolute value of these residuals, and that is exactly what the **Least Absolute Deviations** methodology employs.

$$\min \sum_{i=1}^{N} |e_i|$$

Least Absolute Deviations

That more straightforward manner would just be minimizing the sum of the absolute value of these residuals, and that is exactly what the **Least Absolute Deviations** methodology employs.

$$\min \sum_{i=1}^{N} |e_i|$$

- We refer to this as the *median regression*, and in some ways, it is quite a bit better than looking at things at the mean.

Least Absolute Deviations

That more straightforward manner would just be minimizing the sum of the absolute value of these residuals, and that is exactly what the **Least Absolute Deviations** methodology employs.

$$\min \sum_{i=1}^{N} |e_i|$$

- We refer to this as the *median regression*, and in some ways, it is quite a bit better than looking at things at the mean.
- But the median is just one part of the distribution, what about the other 98 parts of the distribution? What can we do with those?

Quantile Regression

Quantile Regression

The **Quantile Regression** is a non-parametric regression which minimizes a sum of asymmetric penalties for underprediction $q \cdot |e_i|$ and overprediction $(1 - q) \cdot |e_i|$. If you recall, $e_i = y_i - x\beta$

$$\min \left(\sum_{i=1}^{N} q|e_i| + \sum_{i=1}^{N} (1 - q)|e_i| \right)$$

Quantile Regression

The **Quantile Regression** is a non-parametric regression which minimizes a sum of asymmetric penalties for underprediction $q \cdot |e_i|$ and overprediction $(1-q) \cdot |e_i|$. If you recall, $e_i = y_i - x\beta$

$$\min \left(\sum_{i=1}^{N} q|e_i| + \sum_{i=1}^{N} (1-q)|e_i| \right)$$

More formally, the qth quantile regression estimator $\hat{\beta}_q$ minimizes with respect to β_q the objective function

$$Q = \sum_{i \in y_i \geq x\beta}^{N} q|y_i - x_i\beta_q| + \sum_{i \in y_i < x\beta}^{N} (1-q)|y_i - x_i\beta_q|$$

Some Intuition with the Quantile Regression

Some Intuition with the Quantile Regression

1. Obviously, if $q = 0.5$, the quantile regression equation reduces to the LAD.

Some Intuition with the Quantile Regression

1. Obviously, if $q = 0.5$, the quantile regression equation reduces to the LAD.
2. Quantile regressions use linear programming techniques like the simplex method to obtain estimates, it doesn't use OLS or maximum likelihood.

Some Intuition with the Quantile Regression

1. Obviously, if $q = 0.5$, the quantile regression equation reduces to the LAD.
2. Quantile regressions use linear programming techniques like the simplex method to obtain estimates, it doesn't use OLS or maximum likelihood.
3. Because it can't be estimated using OLS, quantile regression cannot be decomposed into an ANOVA framework, just a psudeo one.

Some Intuition with the Quantile Regression

1. Obviously, if $q = 0.5$, the quantile regression equation reduces to the LAD.
2. Quantile regressions use linear programming techniques like the simplex method to obtain estimates, it doesn't use OLS or maximum likelihood.
3. Because it can't be estimated using OLS, quantile regression cannot be decomposed into an ANOVA framework, just a psudeo one.
4. Quantile coefficients must be interpreted at a specific quantile. Hence, you have 99 estimates for a β, one for each quantile.

Advantages of the Quantile Regression

Advantages of the Quantile Regression

- Robust to outliers compared to OLS or typical MLE.

Advantages of the Quantile Regression

- Robust to outliers compared to OLS or typical MLE.
- Flexibility for modeling data, especially for those with non-normal distributions.

Advantages of the Quantile Regression

- Robust to outliers compared to OLS or typical MLE.
- Flexibility for modeling data, especially for those with non-normal distributions.
- Richer characterization and description of the data. Low effects, high effects, the entire spectrum are obtainable.

Let's Look at the LAD Estimate

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

tau: [1] 0.5

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

Let's Look at the LAD Estimate

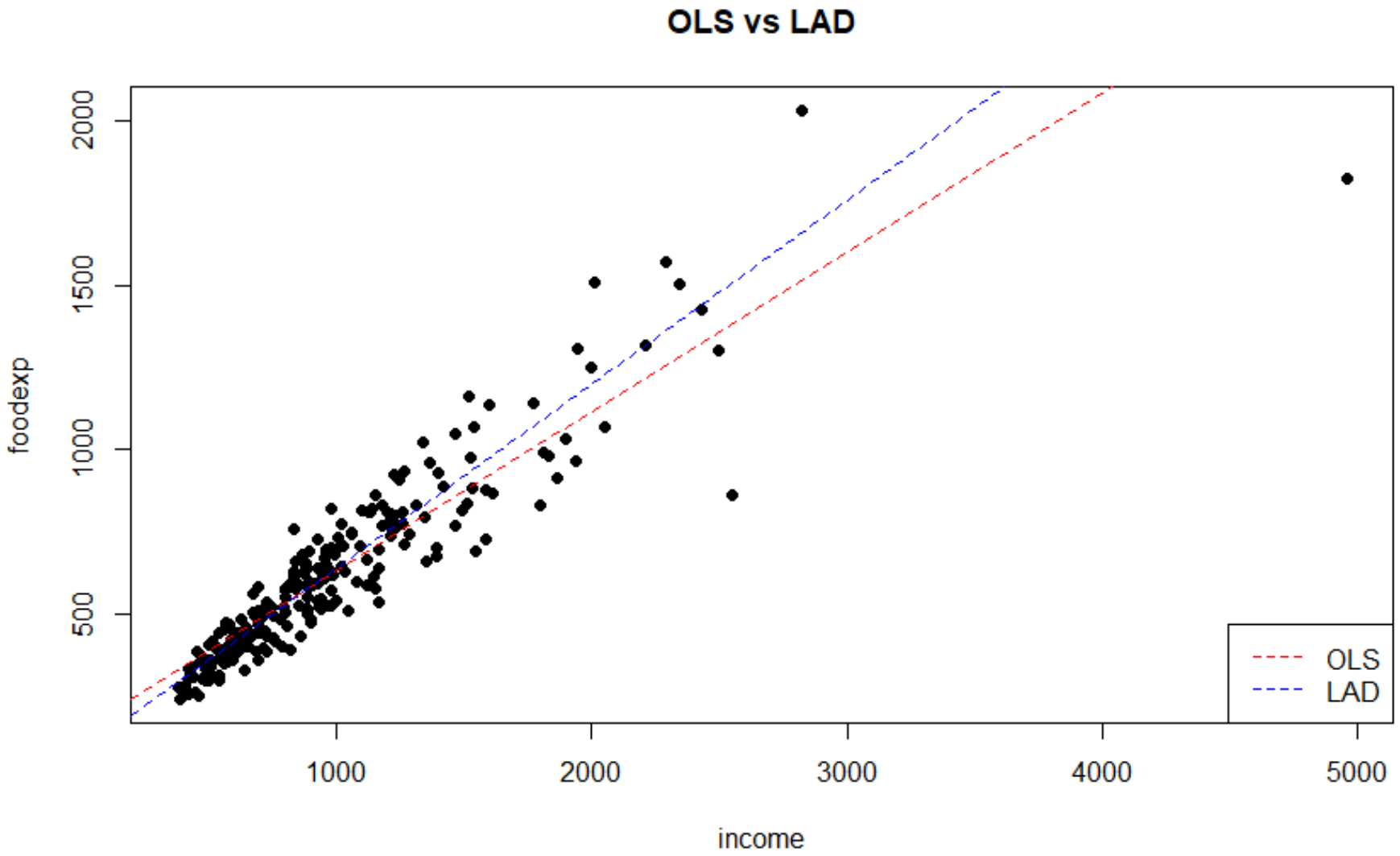
```
Call: rq(formula = foodexp ~ income, tau = 0.5, data = engel)

tau: [1] 0.5

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

We see a different β as compared to β_{OLS}. This could potentially be a better characterization of the data because the estimate is not as affected by outliers. But we can do much more!

OLS vs. LAD


OLS vs LAD
foodexp
500
1000
1500
2000
1000
2000
3000
4000
5000
income
OLS
LAD

Let's Look at the Extremes

Let's Look at the Extremes

Let's see the quantile regression for $q = 0.01$ (i.e. 1st quantile)

```
tau: [1] 0.01

Coefficients:
            coefficients lower bd  upper bd
(Intercept) 131.08192    119.05703 131.69391
income        0.28720      0.28720   0.30454
```

Let's Look at the Extremes

Let's see the quantile regression for $q = 0.01$ (i.e. 1st quantile)

```
tau: [1] 0.01

Coefficients:
            coefficients lower bd  upper bd
(Intercept) 131.08192    119.05703 131.69391
income        0.28720      0.28720   0.30454
```

Let's see the quantile regression for $q = 0.99$ (i.e. 99th quantile)

```
tau: [1] 0.99

Coefficients:
            coefficients lower bd  upper bd
(Intercept)  95.81835     -0.48477 222.06367
income        0.70387      0.65887   1.11211
```

Intuition of the Quantile Results

Food Expenditure	OLS	$q = 10$	$q = 50$ (LAD)	$q = 90$
Income	0.49	0.40	0.56	0.69
Intercept	147.48	110.14	81.48	67.35

- For every dollar increase in income, indviduals spend 0.40 more on food for those with low food expenditures as compred to 0.69 more for those with high food expenditures.
- May seem odd at first (would you have expected different?) but likely just due to the sample.

Types of Significance

Types of Significance

Quantile Coefficients may be tested (individually) using two distinct inferences

Types of Significance

Quantile Coefficients may be tested (individually) using two distinct inferences

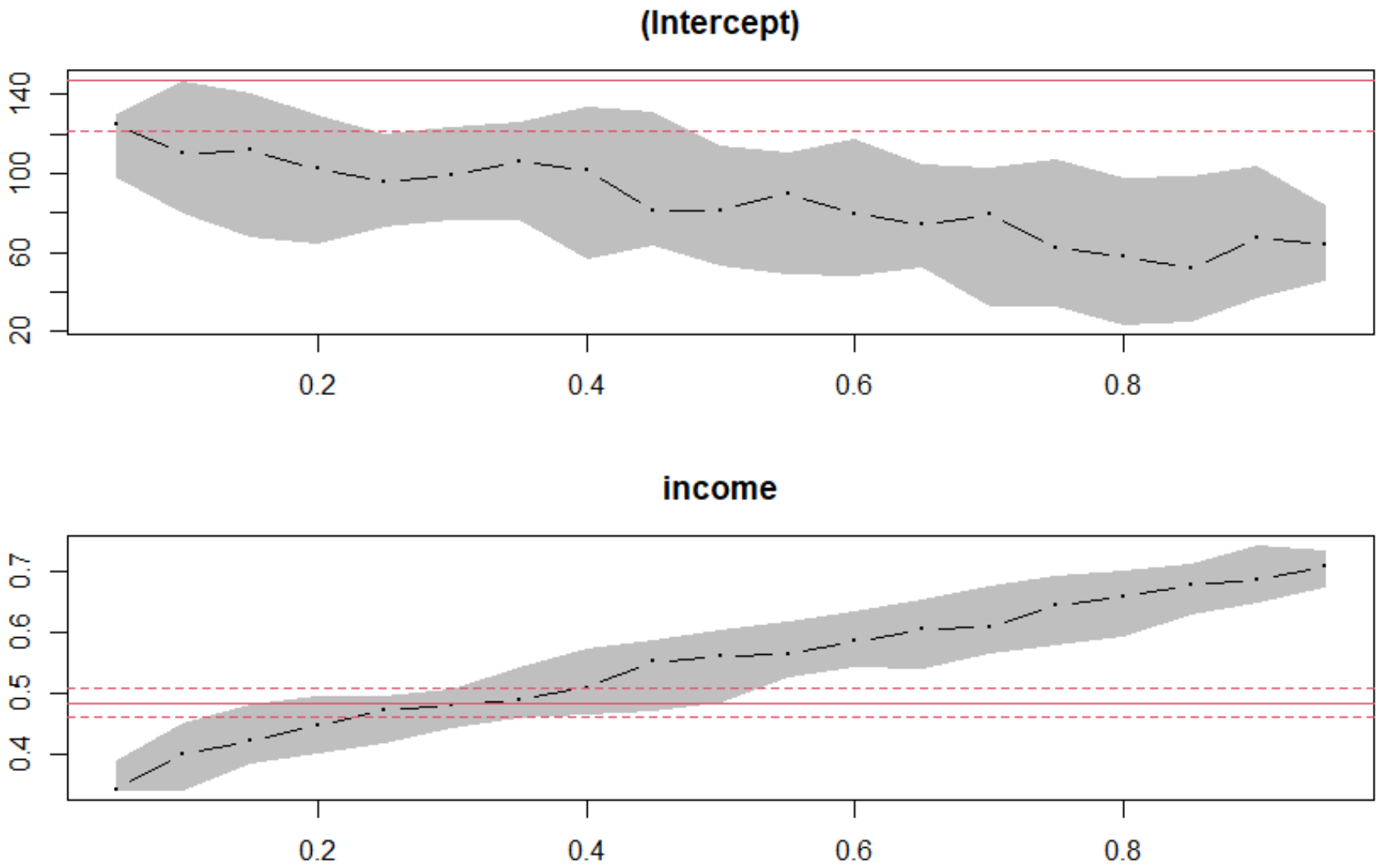
1. Significantly Different from Zero

Types of Significance

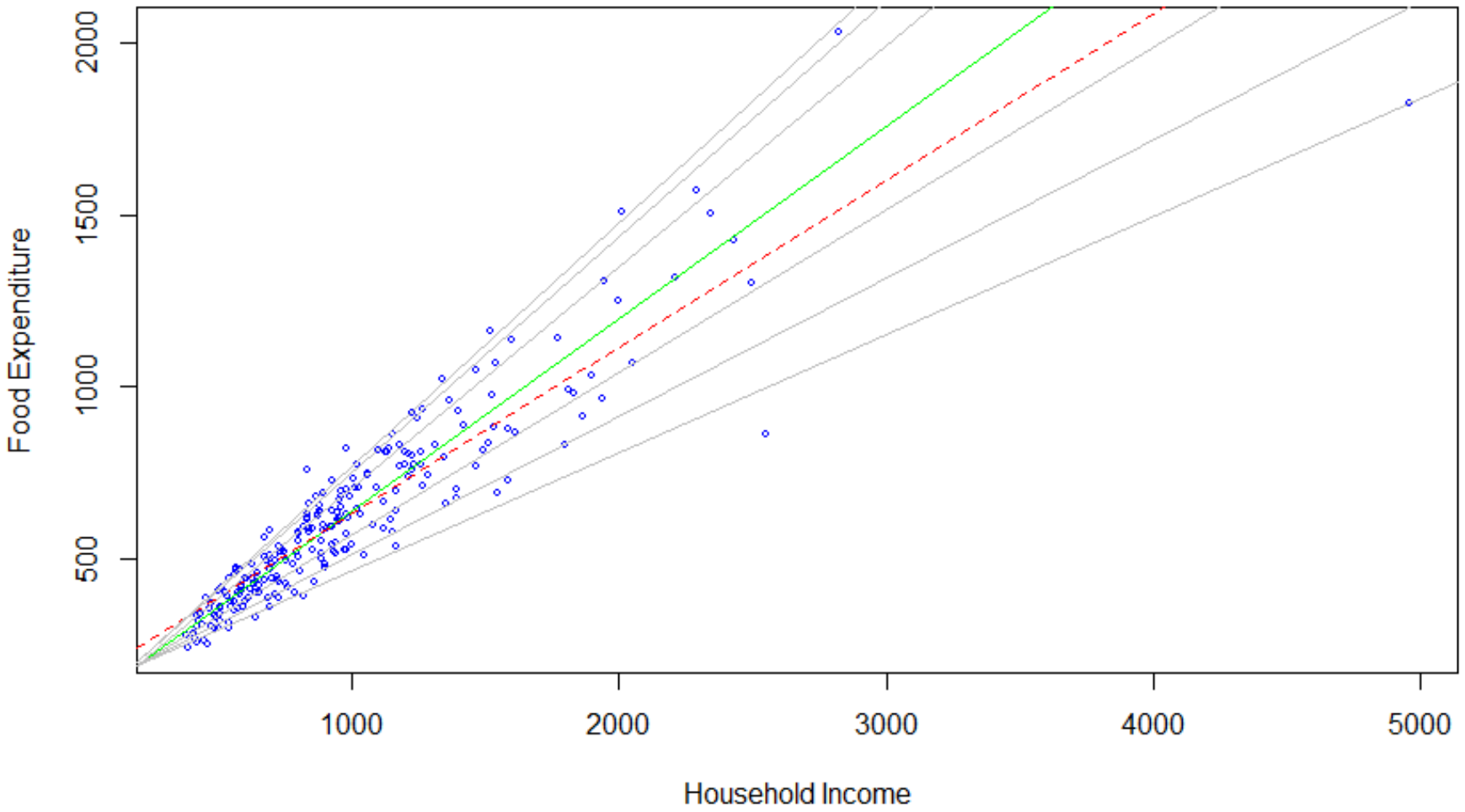
Quantile Coefficients may be tested (individually) using two distinct inferences

1. Significantly Different from Zero
2. Significantly Different from OLS

Visualizing all Coefficients



Bringing it all Together



Let's try another example

Total medical expenditures	OLS regression	Quantile regression at 0.25 quantile	Quantile regression at 0.5 quantile	Quantile regression at 0.75 quantile
Supplementary private insurance	585*	453*	687*	708
Number of chronic problems	2528*	782*$^{+}$	1332*$^{+}$	2855*
Age	7*	16*	35*	87*
Female	-1239	16^{+}	-260^{+}	-554
White	2193	338	632	801
Intercept	461	-1412	-2252*	-4512

*: Significantly different quantile regression coefficient from zero at the 5% significance level.

$^{+}$: Significantly different quantile regression coefficients from OLS coefficients at the 5% significance level, when the OLS coefficient is outside of the quantile regression coefficient confidence interval.

Source: Ani Katchova

Sources

- Katchova, A. (2013) Quantile Regression
- Koenker, R. (2021) Quantile Regression in R: A Vignette
- Koenker, R., & Bassett Jr, G. (1978). Regression quantiles. Econometrica: journal of the Econometric Society, 33-50.