Why Quantile Regression?

Quantile Regression

Session 4.1 Lecture Video Slides

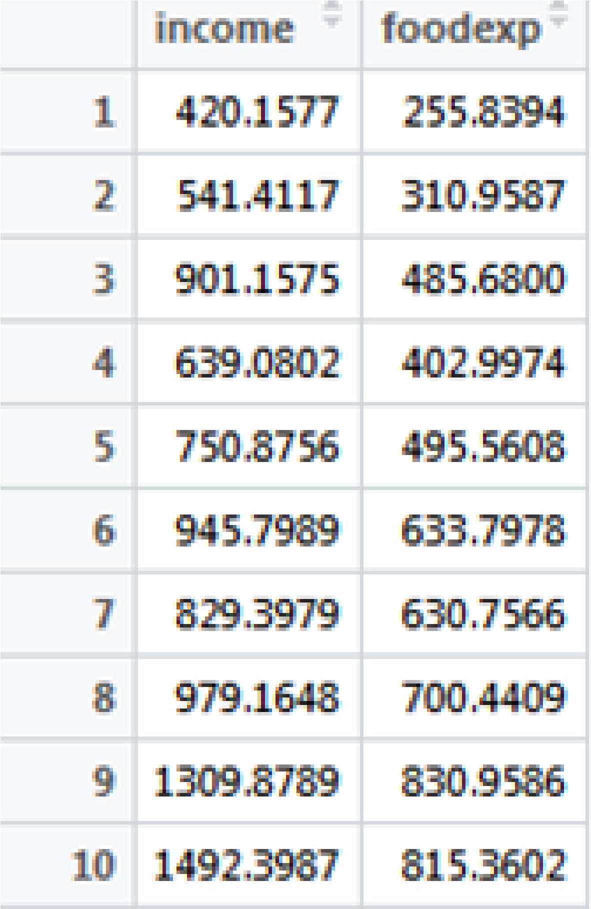
Background

- 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 Diagnostic Plots in R with plot().
- What if one or more assumptions are not satisfied?
- Knowing only linear regression, we can try to do mathematical transformations of the variables and then 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 not met.

Engel Dataset from Rpackage quantreg

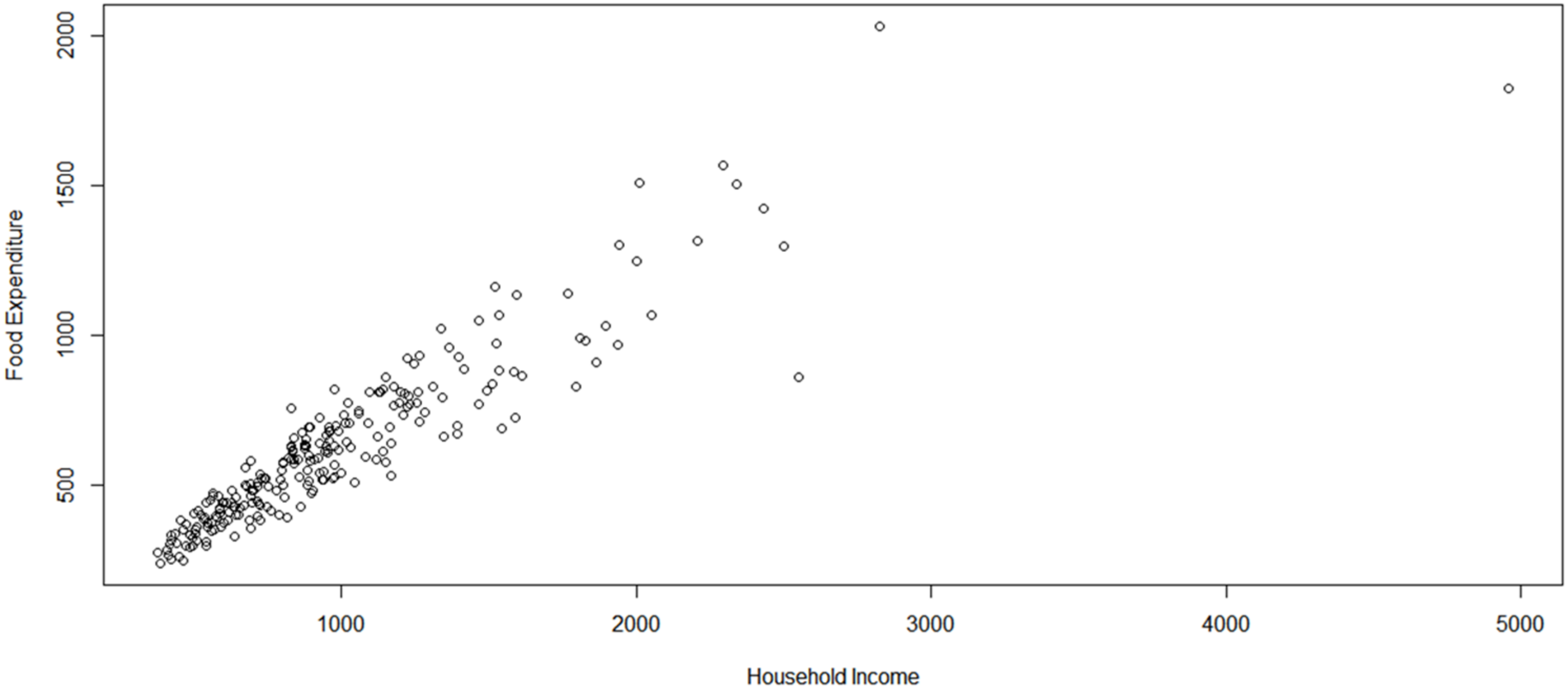
- Dataset that records Family Expenditure on Food and Family Income in Belgium 1857.
- Used to show a limitation of Linear Regression and usefulness of Quantile Regression.
- Dataset is in quantreg Rpackage.



	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.

Scatterplot of Food Expenditure vs Household Income



- What is the business purpose of analyzing this data?
- To study how an essential cost of living (food) varies as income varies.

Review of Linear Regression Model

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

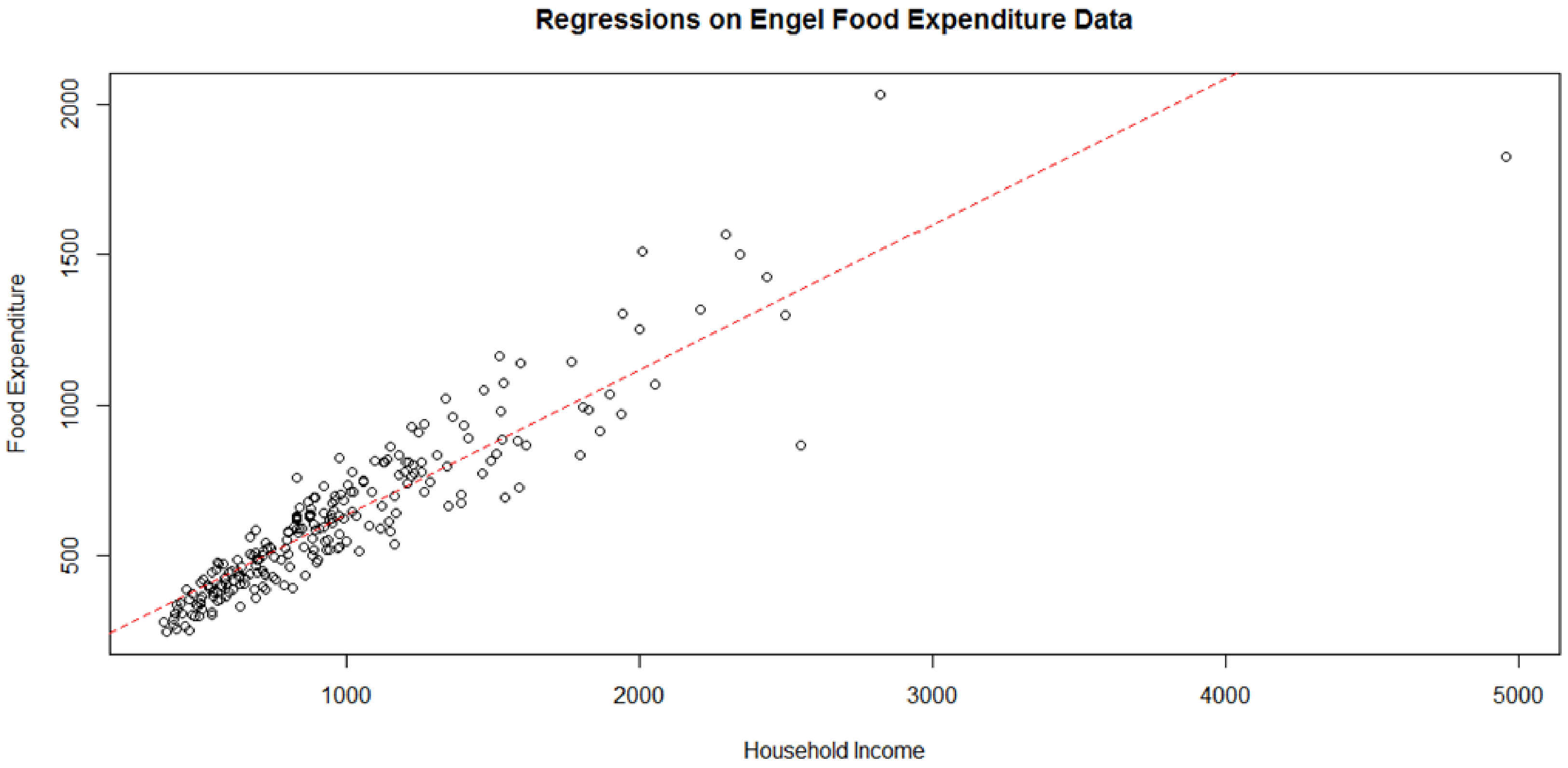
Straight Line Equation

e ~ N(0, σ)

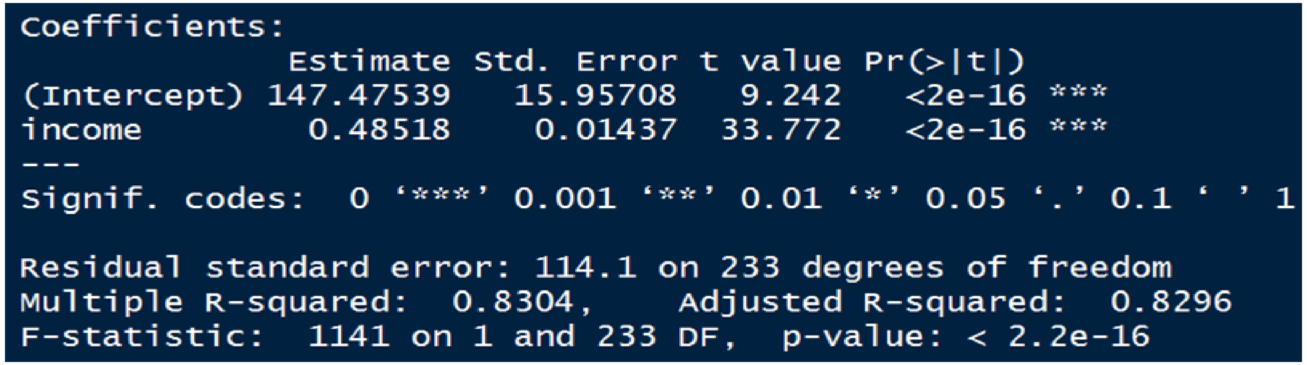
Errors (aka Residuals) follow a Normal Distribution with mean 0 and constant standard deviation.

Q: What does the straight line equation actually represent?

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



Regressions on Engel Food Expenditure Data

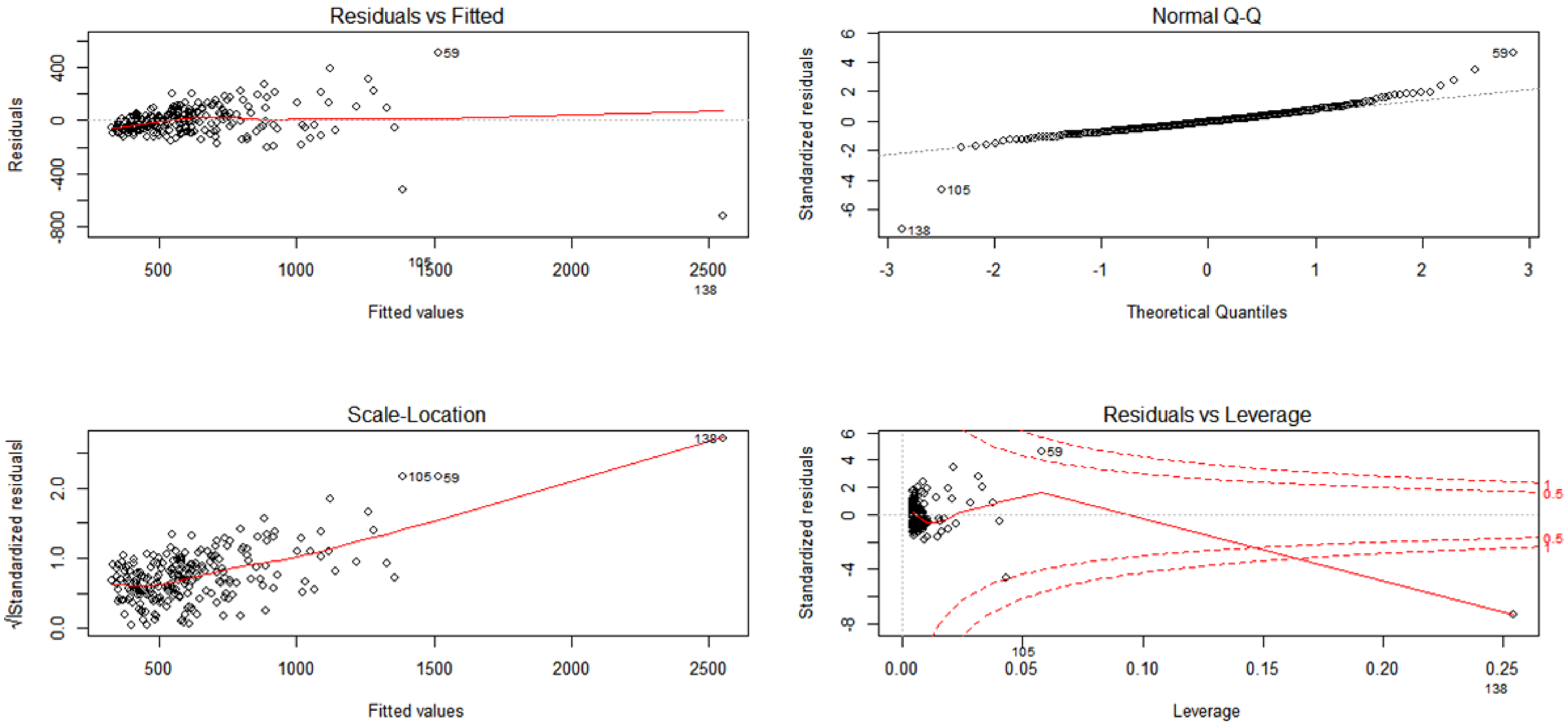


```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 147.47539   15.95708   9.242   <2e-16 ***
income        0.48518    0.01437  33.772   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

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 Obvious Issues



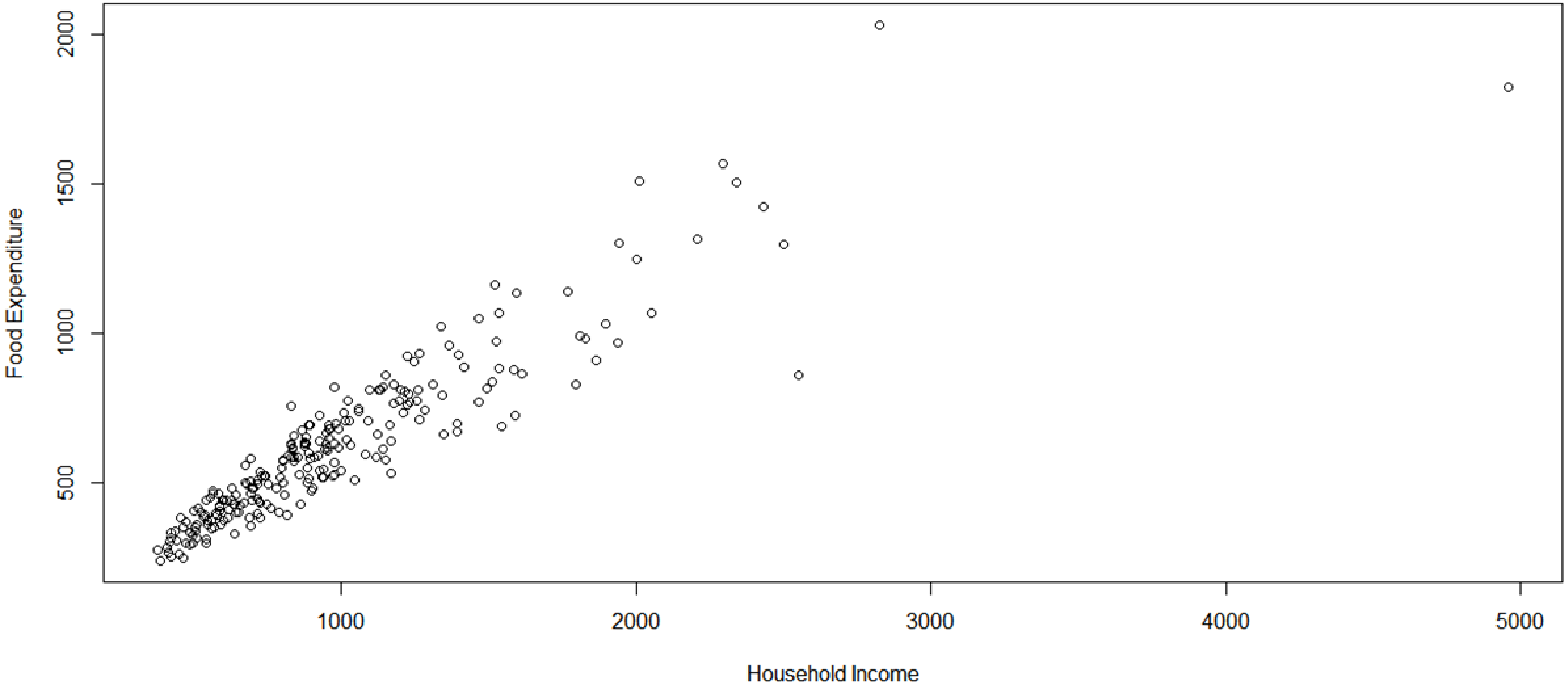
Error Variance increases at higher predicted expenditures.

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/algorithms.)

Specific Business/Social Questions to be Answered?

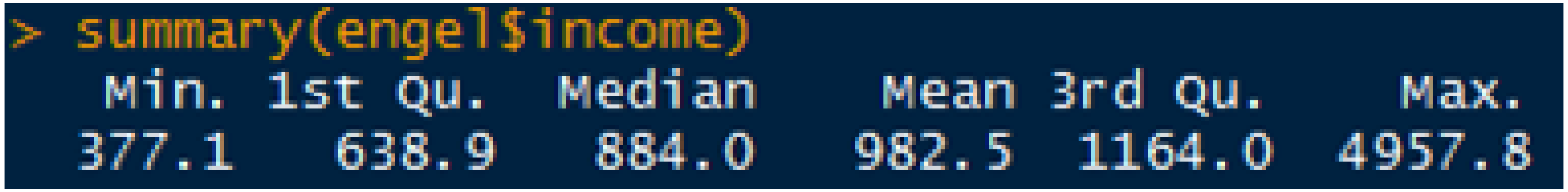


How much does a typical family spend on Food?

Critical Thinking on the Business/Social Purpose

How much does a typical family spend on Food?

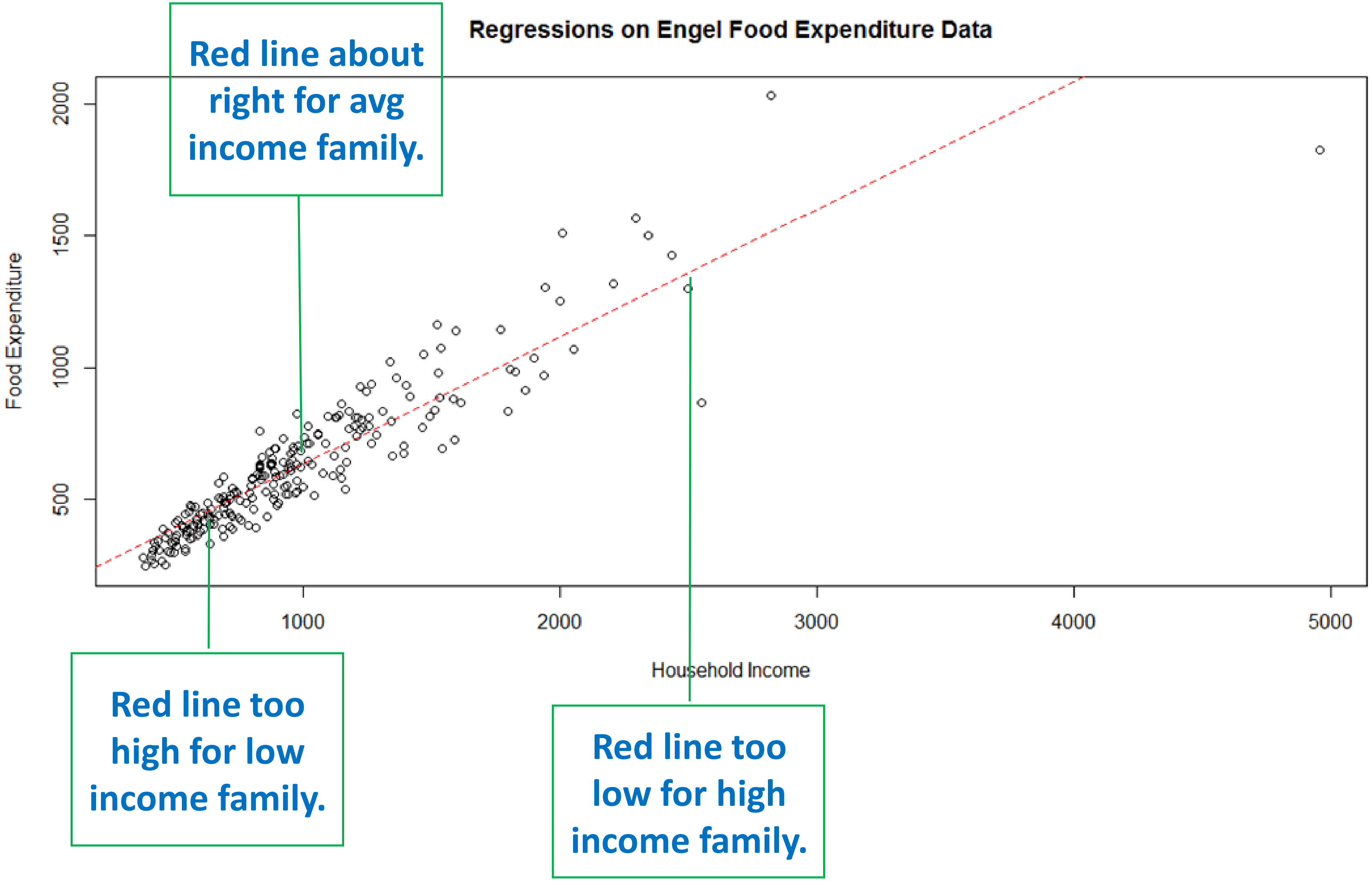
- What do you mean by typical family?
  - Family with mean income?
  - Is this the only kind of family that one is interested in analysing?



```
> 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 expenditures.



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 is often inadequate in such cases.

Next Video

- Rpackage quantreg.
- Original Source from Prof. Roger Koenker.
  - Popularized Quantile Reg to different disciplines.
  - Wrote the Rcode.
  - Prepared the data.
- Python libraries.