Quantile Regression

Ass.Prof. Andriy Stavytskyy

References

- **Buchinsky M.** (1998), “Recent Advances in Quantile Regression Models”, Journal of Human Resources, Vo. 33, Pps. 88-126.
- **Koenker R**. (2005) “Introduction to Quantile Regression”, Econometric Society Monograph Series, Cambridge University Press.

Agenda

- Motivation of Quantile Regression
- Quantile regression Estimation
- Properties of the Estimator
- Example

MOTIVATION OF QUANTILE REGRESSION

Problems – 1

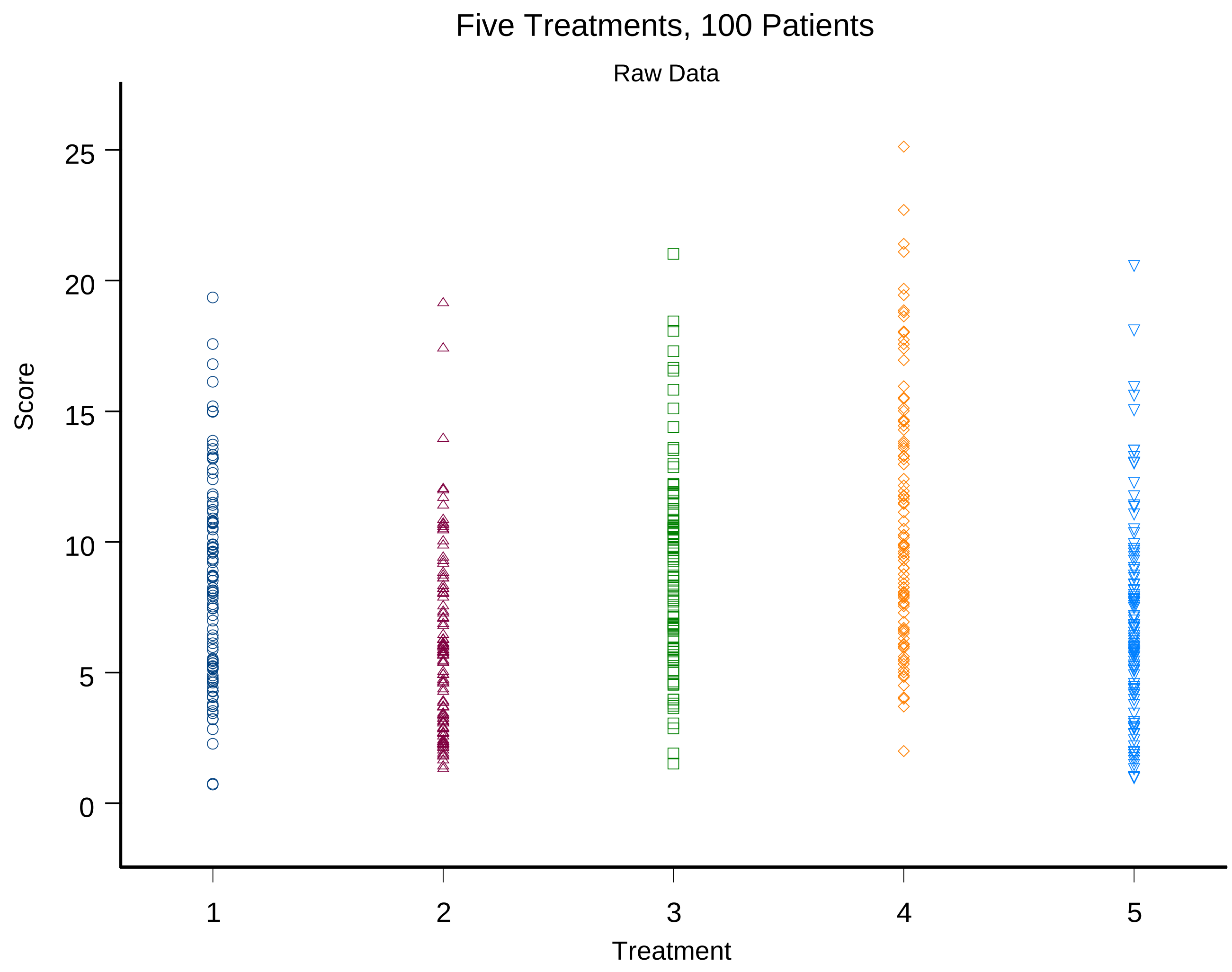
- The distribution of Y, the "*dependent*" variable, conditional on the covariate X, may have ***thick tails***.
- The conditional distribution of Y may be ***asymmetric***.

- The conditional distributio of Y may ***not be unimodal***.

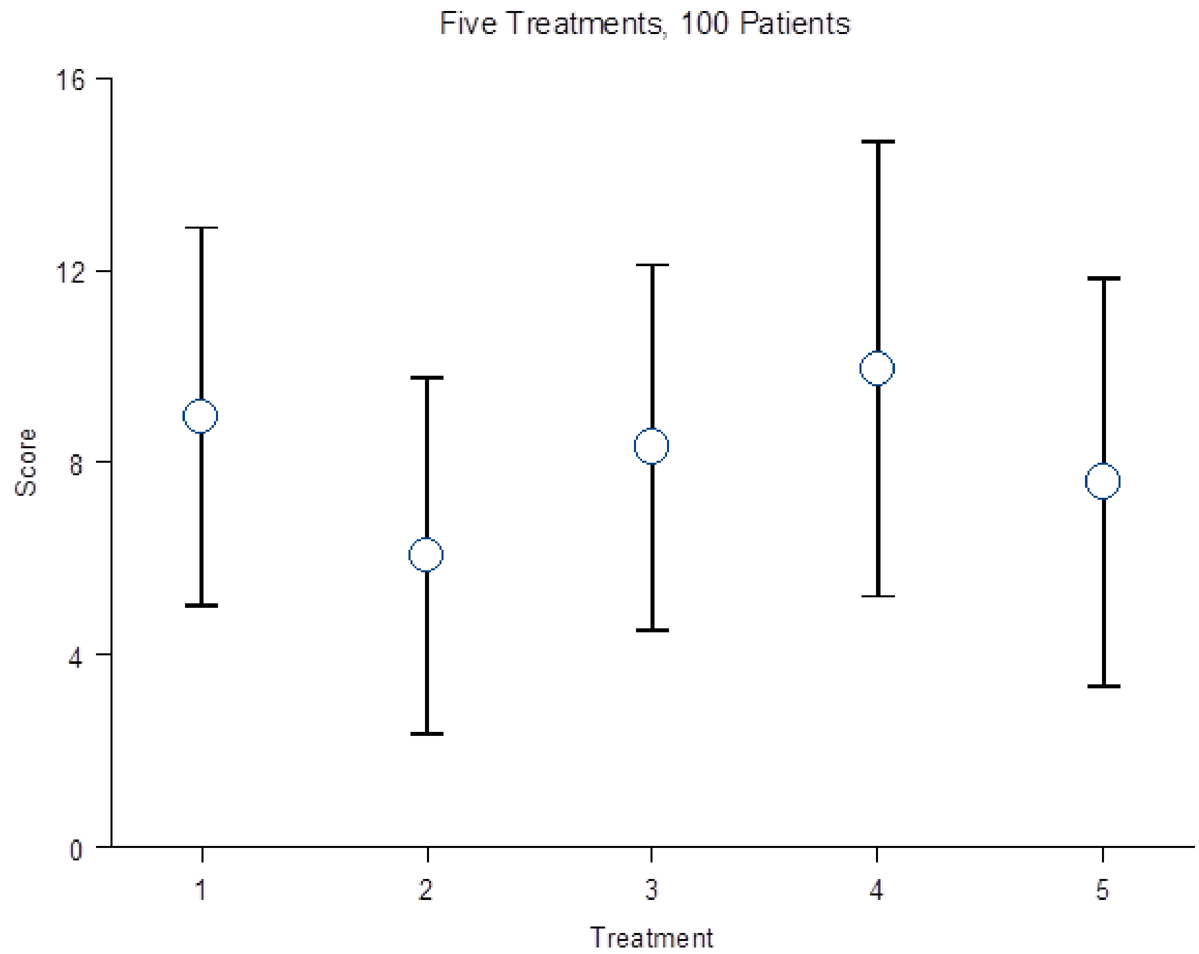
Problems – 2

- ANOVA and regression provide information ***only about the conditional mean***.
- Neither regression nor ANOVA will give us ***robust results***. Outliers are problematic, the mean is pulled toward the skewed tail, multiple modes will not be revealed.
- More knowledge about the distribution of the statistic ***may be important***.
- The covariates may shift not only the location or scale of the distribution, ***they may affect the shape as well.***

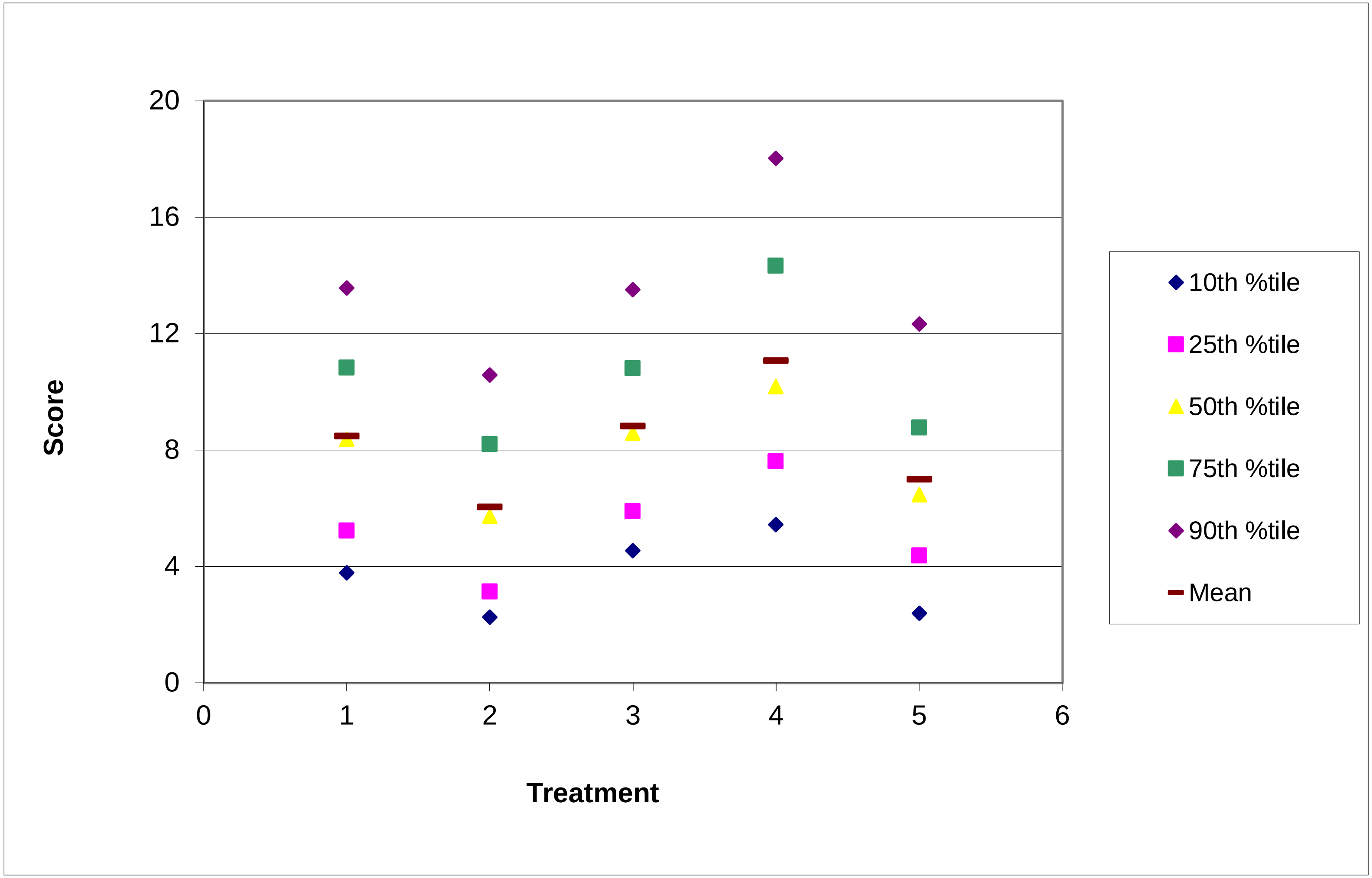
Example: data



Example: means with error bars



Example: quantiles



Reasons to use quantiles rather than means

- Analysis of distribution rather than average
- Robustness
- Skewed data
- Interested in representative value
- Interested in tails of distribution
- Unequal variation of samples

- **E.g.** Income distribution is highly skewed so median relates more to typical person that mean.

QUANTILE REGRESSION ESTIMATION

Quadratic loss function

$$y_t = \beta_1 + \beta_2 x_{2t} + \ldots + \beta_k x_{kt} + u_t,$$

- Ordinarily we specify a quadratic loss function:

$$L(u) = \Sigma u^2$$

- Under quadratic loss we use the conditional mean, via regression or ANOVA, as our predictor of Y for a given X=x.

Quantile definition

- For a given $p \in [0, 1]$ a p^{th} quantile of a random variable Z is any number ζ_p such that

$$\Pr(Z < \zeta_p) \le p \le \Pr(Z \le \zeta_p).$$

- The solution always exists, but needs not be unique.
- Ex: Suppose Z={3, 4, 7, 9, 9, 11, 17, 21} and p=0.5 then

$$\Pr(Z<9) = 3/8 \le 1/2 \le \Pr(Z \le 9) = 5/8$$

Quantiles

Quantiles can be used to characterize a distribution:

- Median
- Interquartile Range
- Interdecile Range
- Symmetry = $(\zeta_{.75} - \zeta_{.5})/(\zeta_{.5} - \zeta_{.25})$
- Tail Weight = $(\zeta_{.90} - \zeta_{.10})/(\zeta_{.75} - \zeta_{.25})$

Quantile Function

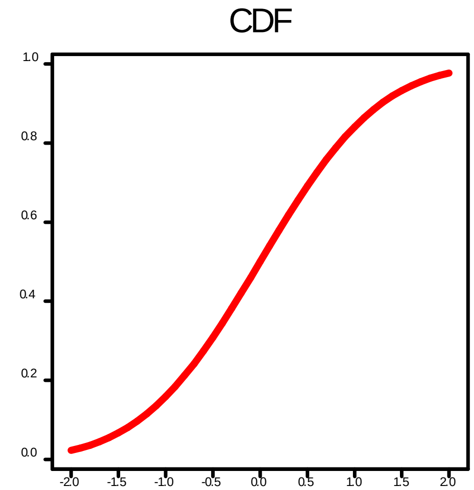
- Cumulative Distribution Function

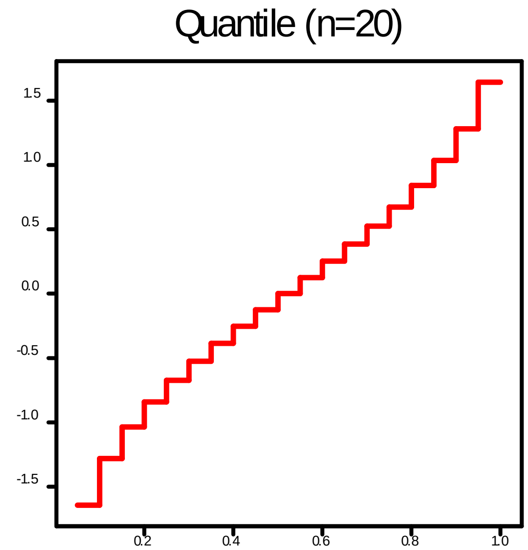
$$F(y) = \text{Prob}(Y \le y)$$

- Quantile Function

$$Q(\tau) = \min(y : F(y) \le \tau)$$

- Discrete step function





Quantile

- Suppose Z is a continuous random variable with cumulative distribution function F(.), then

$$Pr(Z<z) = Pr(Z\leq z)=F(z)$$

for every z in the support and a p^{th} quantile is any number ζ_p such that $F(\zeta_p) = p$

- If F is continuous and strictly increasing then the inverse exists and $\zeta_p = F^{-1}(p)$

The asymmetric absolute loss function – 1

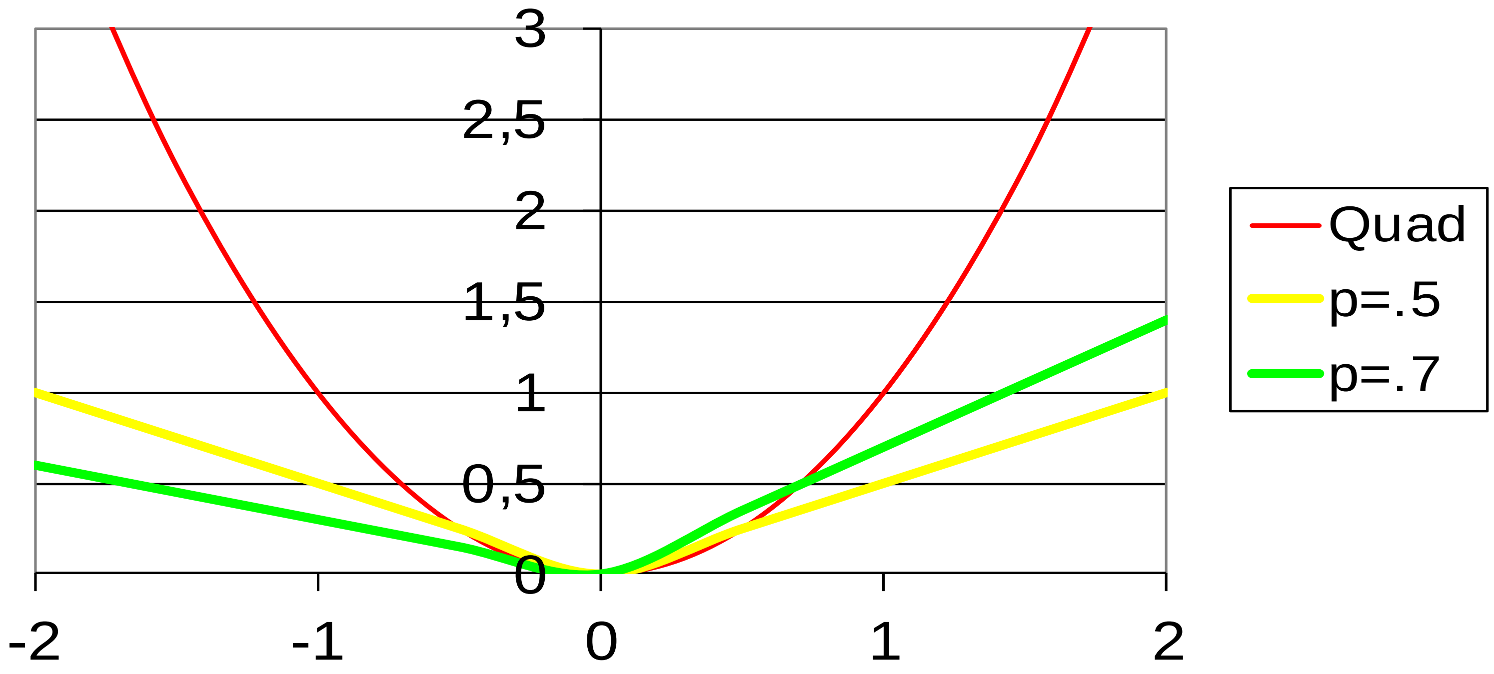
- The asymmetric absolute loss function is

$$L_p = [p\,I(u \geq 0) + (1-p)\,I(u < 0)]|u|$$
$$= [p - I(u < 0)]u$$

where u is the prediction error we have made and I(u) is an indicator function of the sort

$$I(u \geq 0) = \begin{cases} 1, & \text{if } u \geq 0, \\ 0, & \text{if } u < 0. \end{cases}$$

Absolute Loss vs. Quadratic Loss



The asymmetric absolute loss function – 2

- Under the asymmetric absolute loss function L_p a best predictor of Y given X=x is a p^{th} conditional quantile.

$$\zeta_p(x)$$

- For example, if p=.5 then the best predictor is the median.

Simple Quantile Regression – 1

- A parametric quantile regression model is correctly specified if, for example,

$$\zeta_p(x) = q(x, \theta) = \alpha + \beta x$$

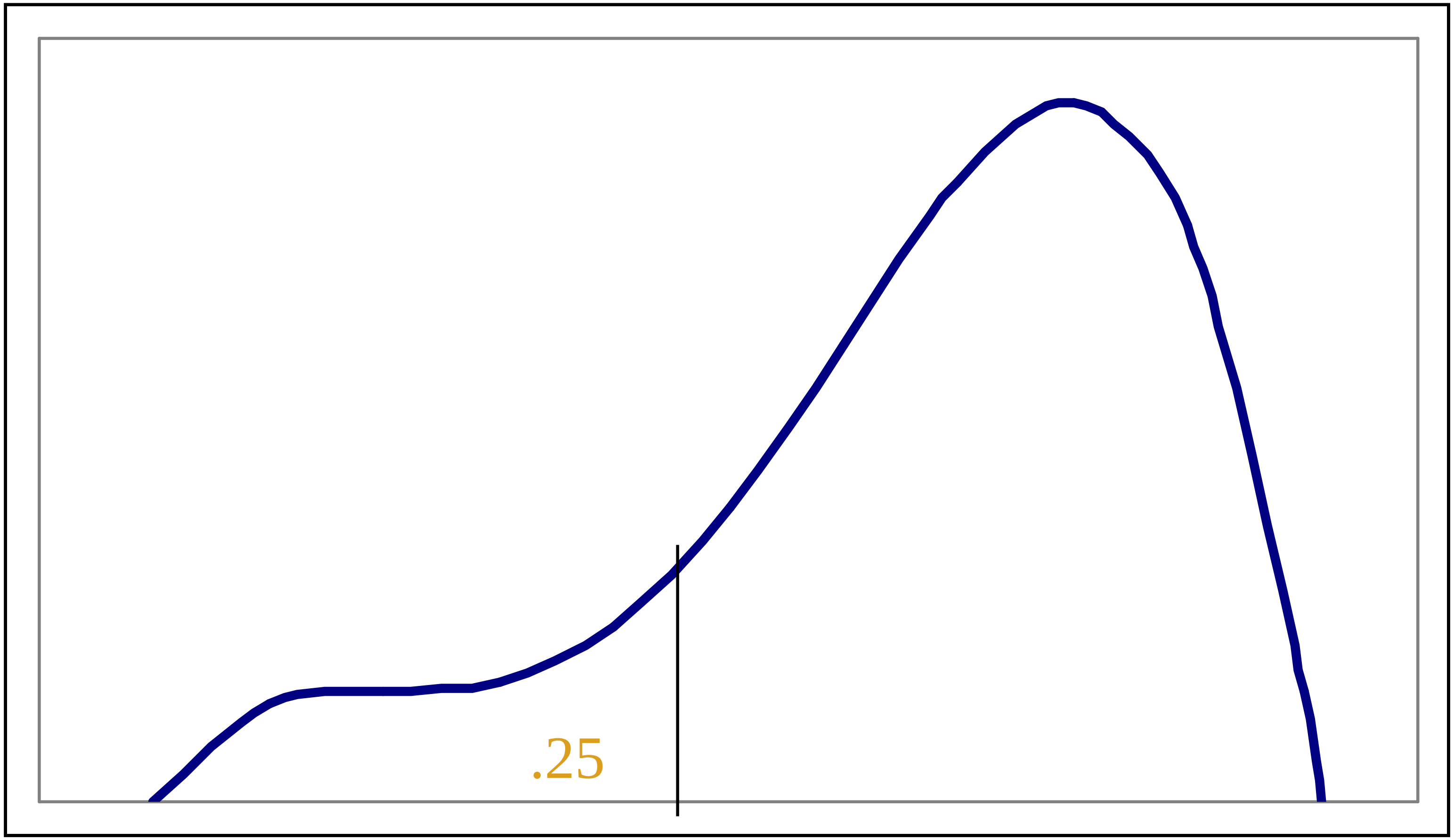
- That is, $\alpha + \beta x$ is a particular linear combination of the independent variable(s) such that

$$p = \Pr(Y \le \zeta_p(x) \mid X = x) = F(\zeta_p(x) \mid x)$$

$$= \Pr(Y \le \alpha + \beta x) = F(\zeta_p(x) - \alpha - \beta x)$$

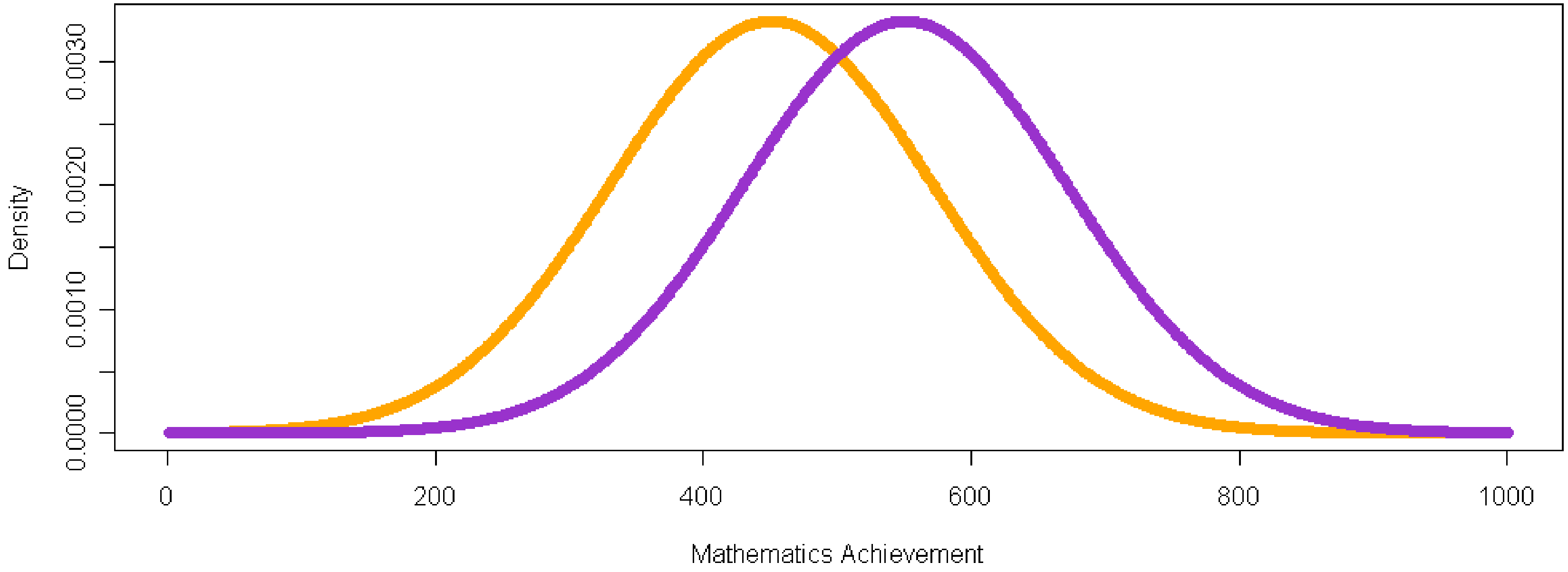
where F( ) is some univariate distribution.

Simple Quantile Regression – 2

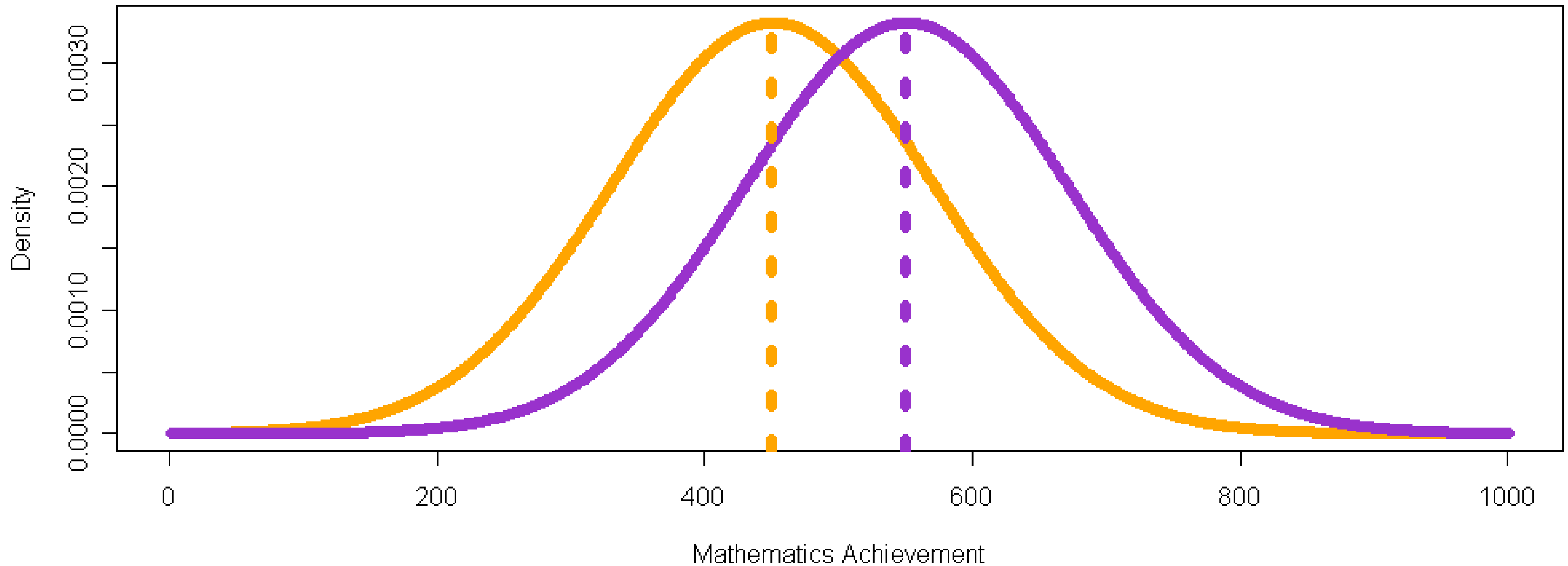


$$\zeta_{.25}(x) = \alpha + \beta x$$

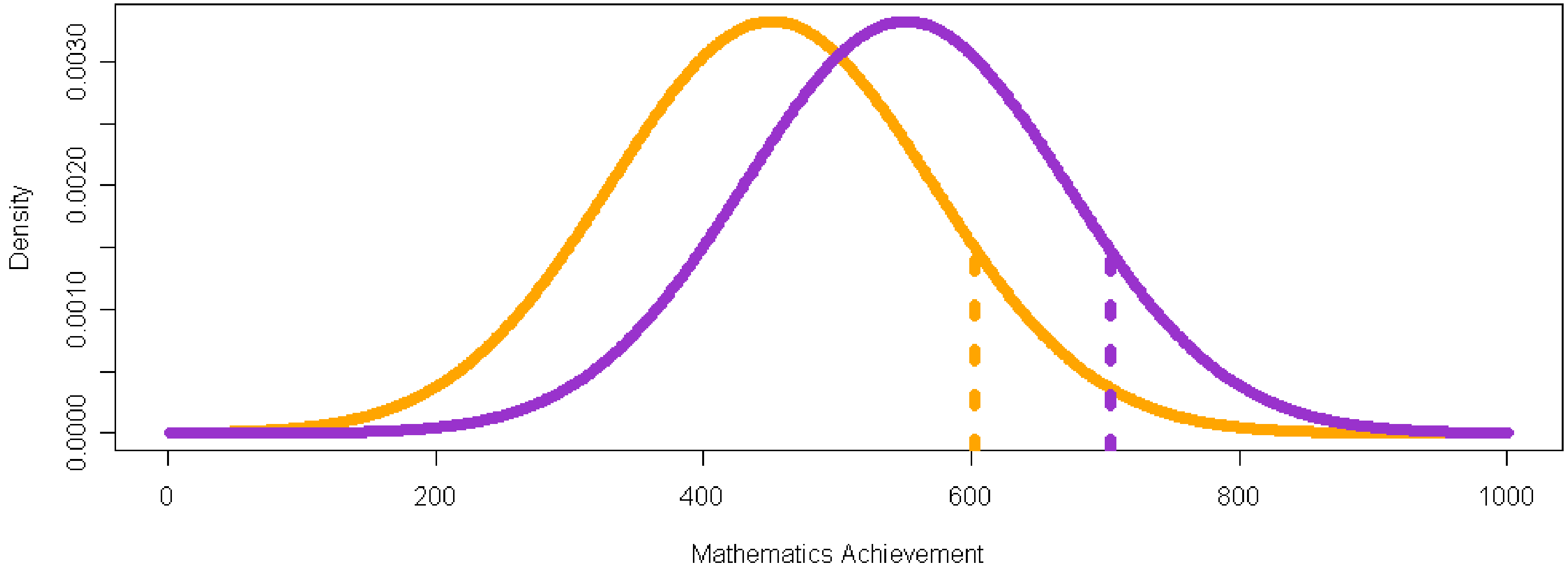
Example: Hypothetical Distributions



Example: OLS Regression Results



Example: Quantile Regression Results

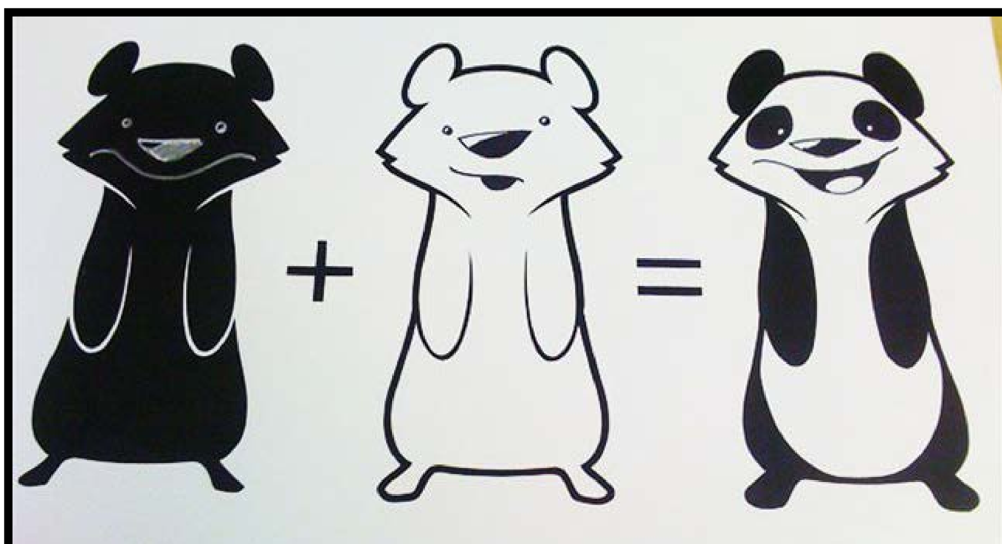


Simple Quantile Regression – 3

A quantile regression model is identifiable if

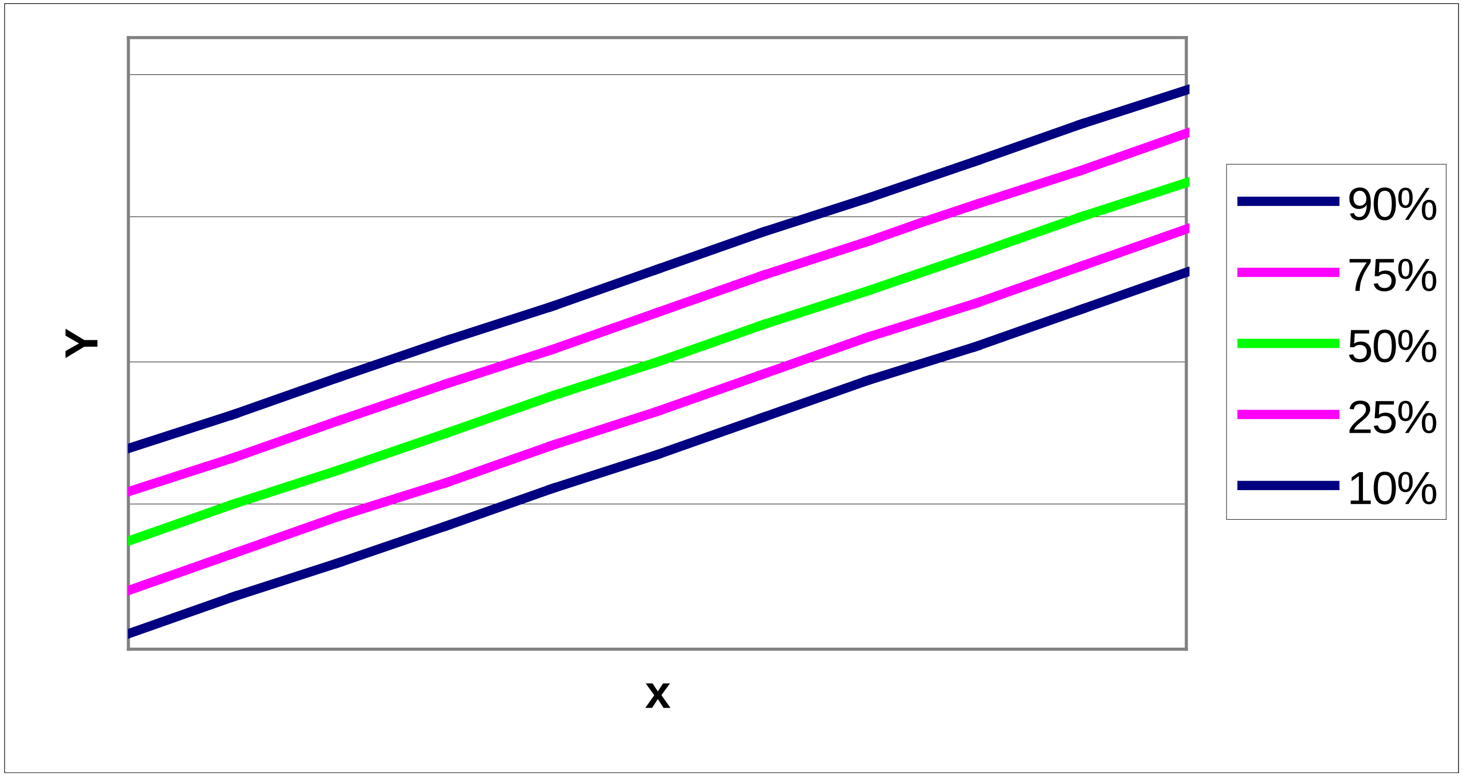
$$\min_{\alpha,\beta} E_F L_p(Y - \alpha - \beta x)$$

has a unique solution.



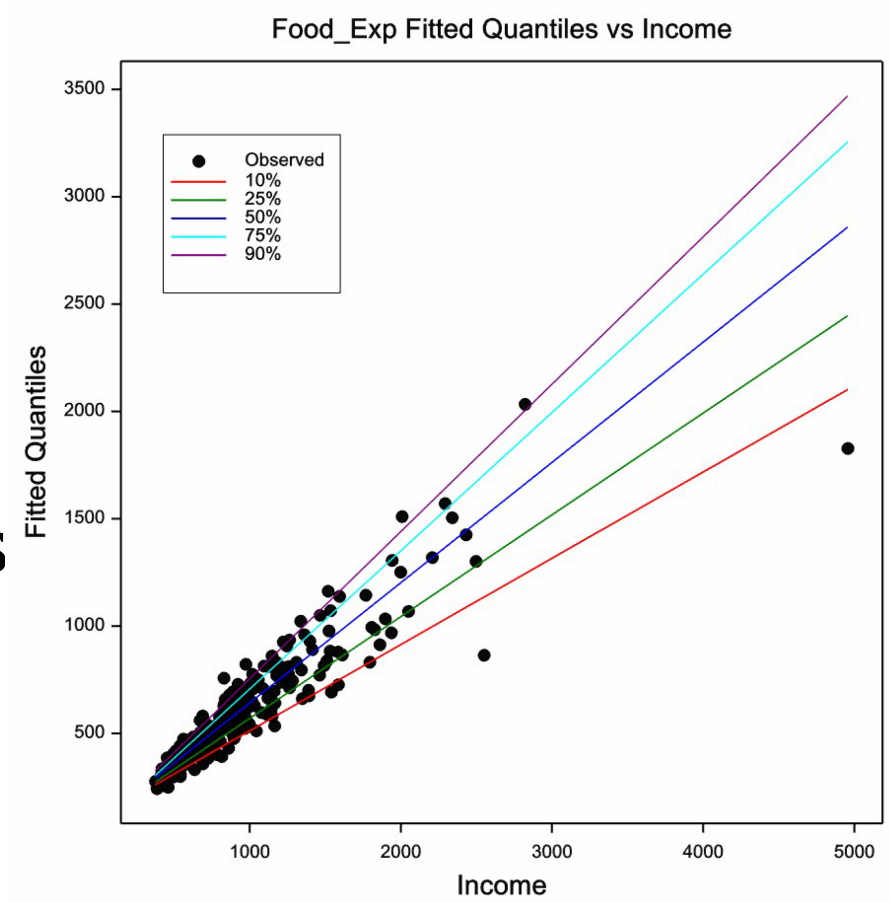
Simple Quantile Regression – 4

- Let $Y=\alpha+\beta x+u$ with $\alpha=\beta=1$, $u \sim N(0,1)$.

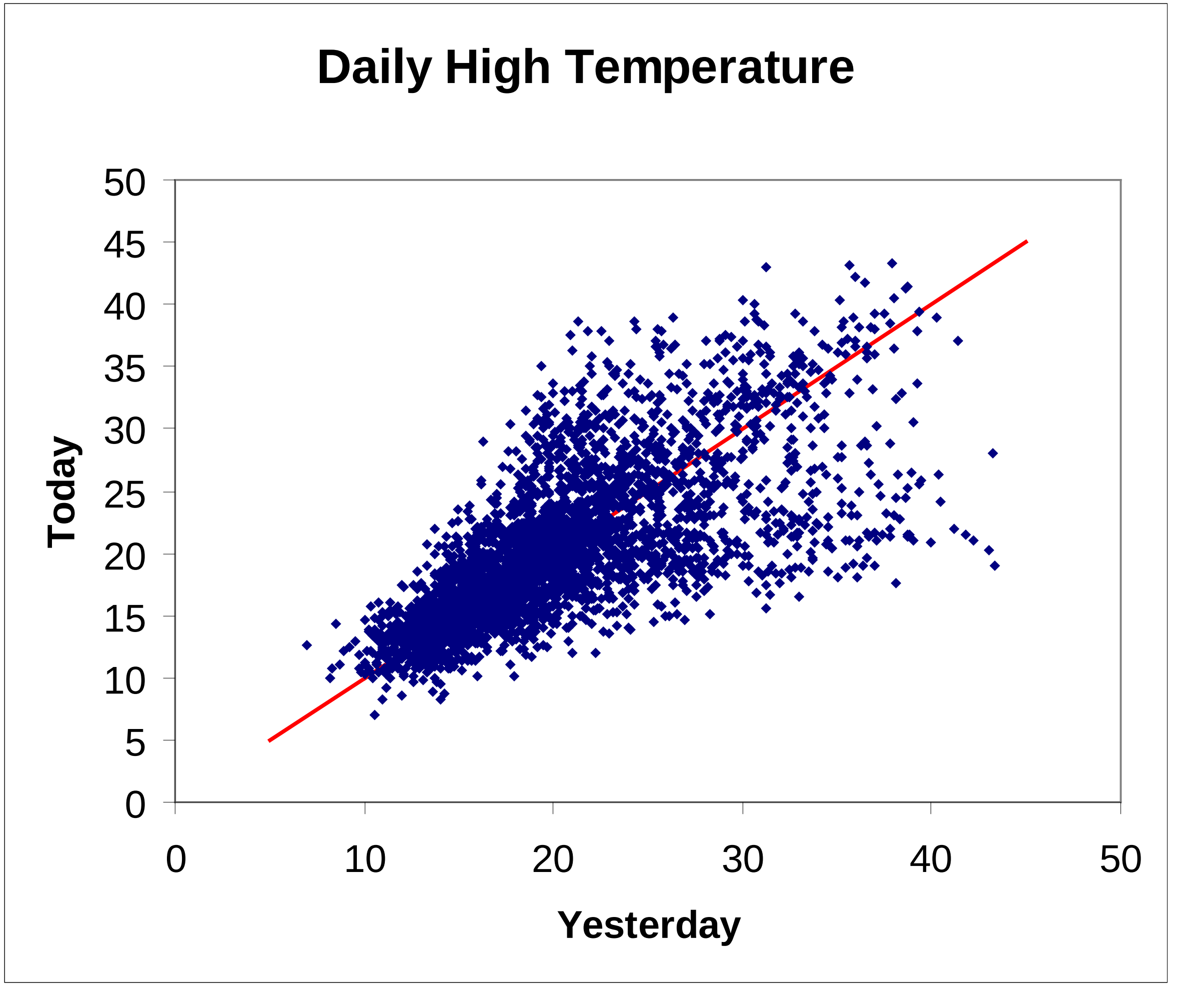


Example: Simple Linear Regression

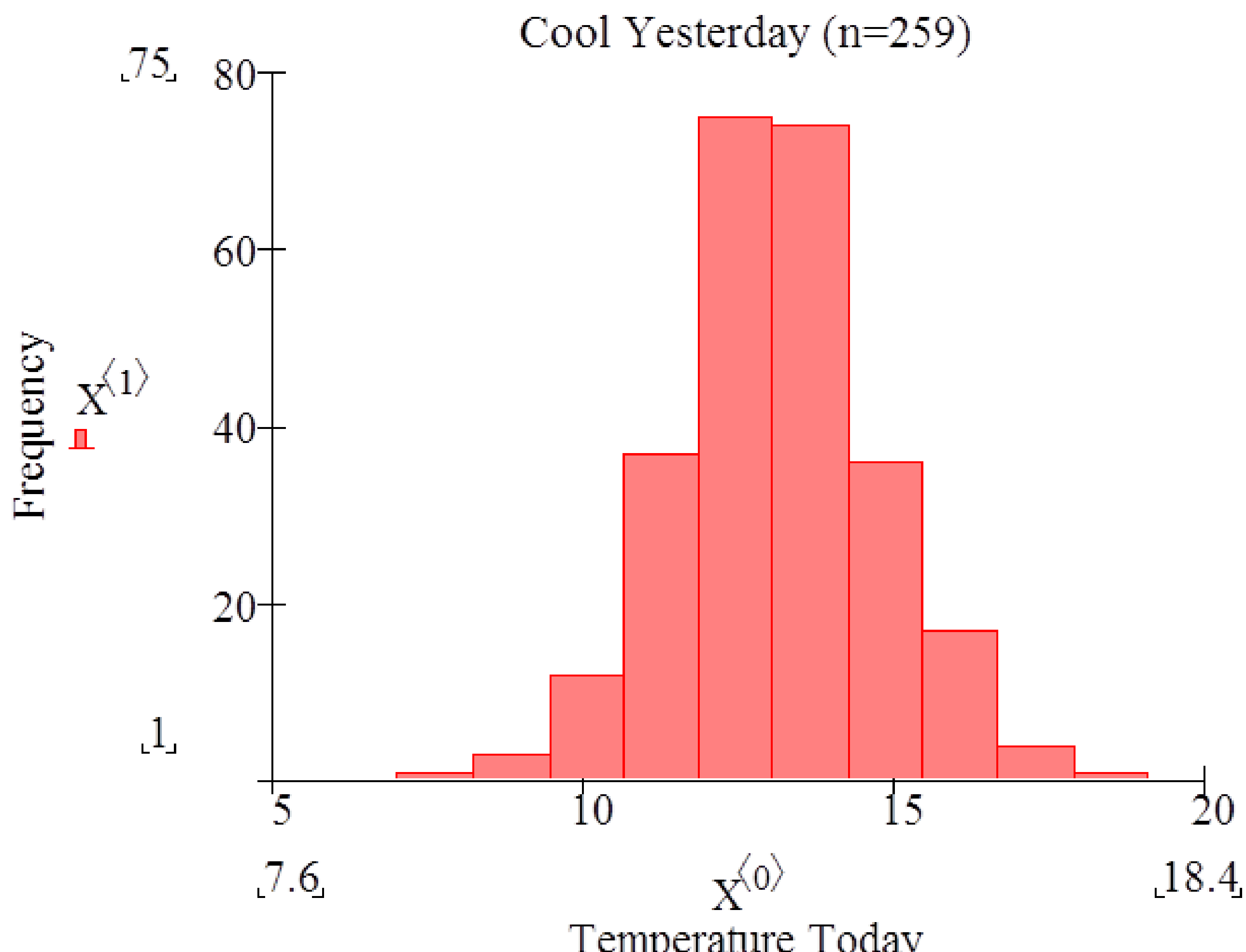
- Food Expenditure **vs** Income
- Engel's (1857) survey of 235 Belgian households
- Change of slope at different quantiles?



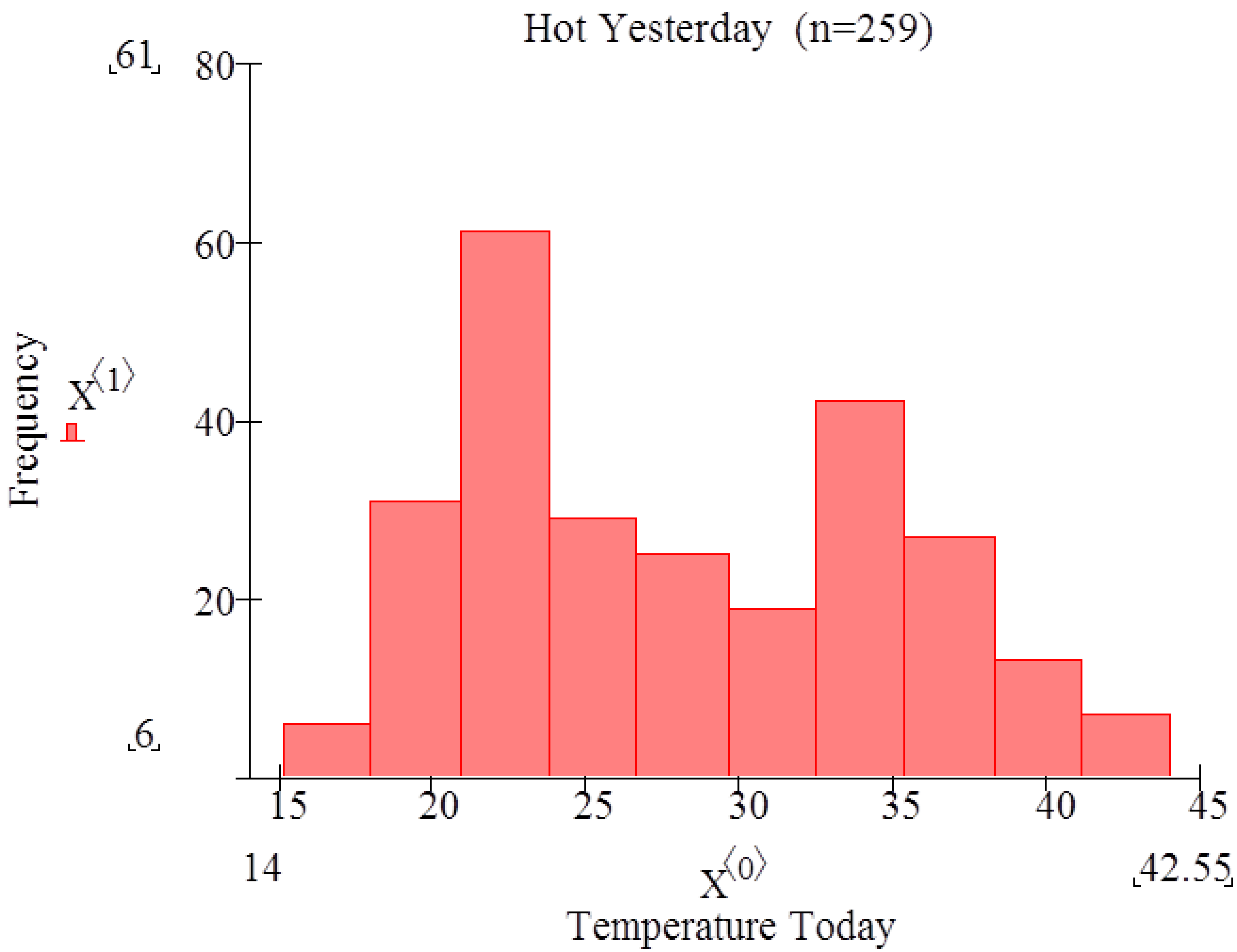
Example: Quantile Regression Analysis – 1



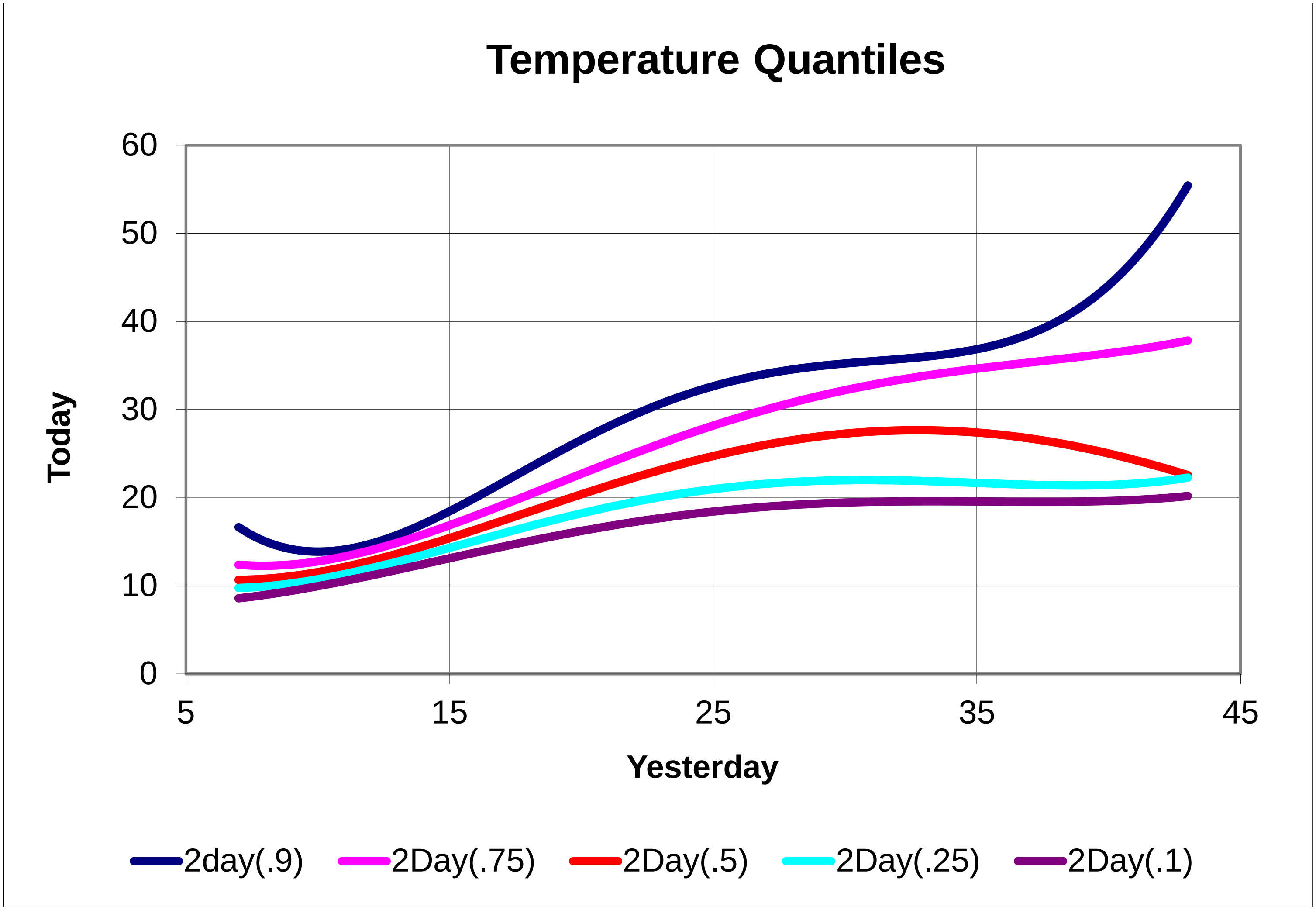
Example: Quantile Regression Analysis – 2



Example: Quantile Regression Analysis – 3



Example: Quantile Regression Analysis – 4



General Quantile Regression

$$y_t = \beta_1 + \beta_2 x_{2t} + \ldots + \beta_k x_{kt} + u_t,$$

$$Y = X\beta + u$$

$$y = X_i \beta + \varepsilon$$

Quantile Regression Estimation – 1

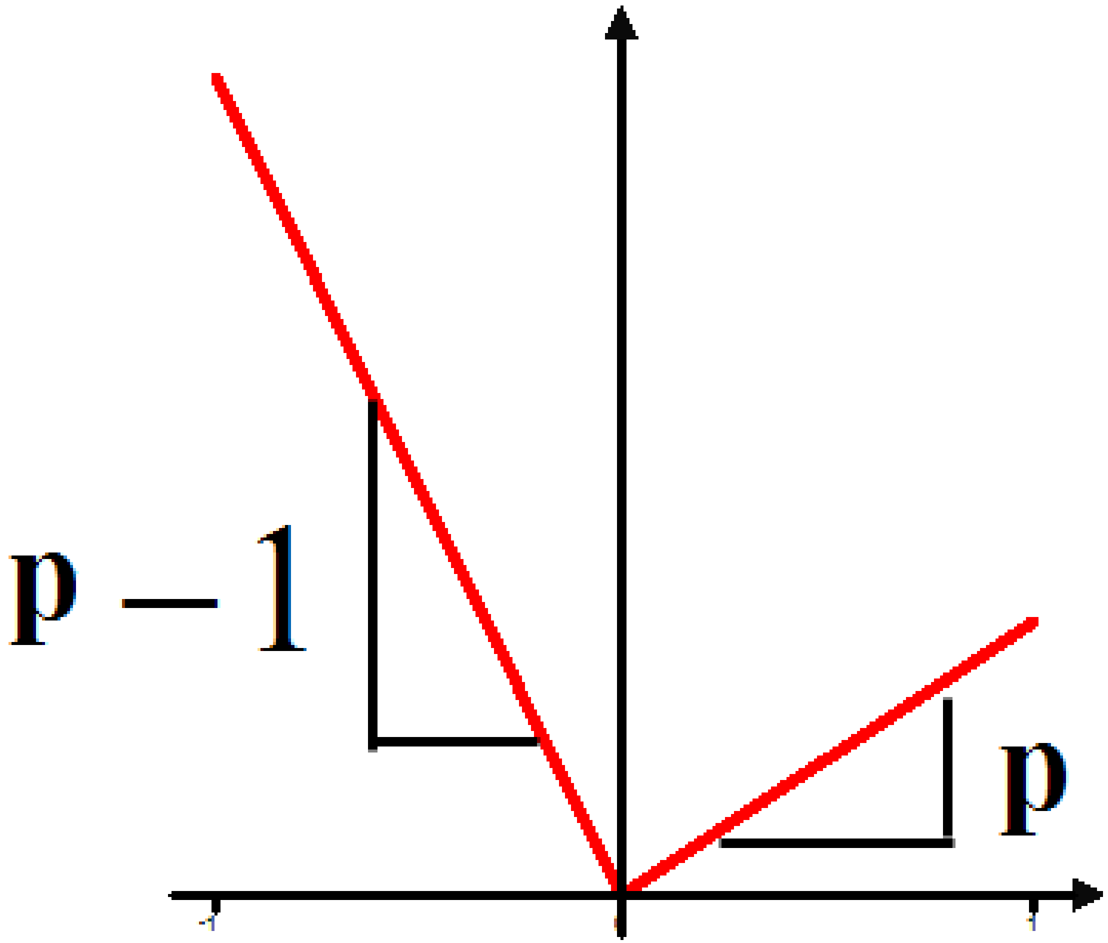
- The quantile regression coefficients are the solution to

$$\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} \left[p - \tfrac{1}{2} - \tfrac{1}{2} \operatorname{sgn}\left(y_i - x_i^T \beta \right) \right] \left(y_i - x_i^T \beta \right)$$

$$\min_{\beta} \left[\sum_{\{i | y_i \geq X_i \beta\}} p \left| y_i - X_i \beta \right| + \sum_{\{i | y_i < X_i \beta\}} (1 - p) \left| y_i - X_i \beta \right| \right]$$

Negative residuals Positive residuals

Quantile Regression Estimation – 2



Quantile Regression Estimation – 3

The k first order conditions are

$$\frac{1}{n}\sum_{i=1}^{n}\left[p-\frac{1}{2}+\frac{1}{2}\operatorname{sgn}\left(y_i - x_i^T\hat{\beta}_p\right)\right]x_i = 0$$

Quantile Regression Estimation – 4

- The fitted line will go through **k** data points.
- The # of negative residuals ≤ **np** ≤ # of neg residuals + # of zero residuals
- The computational algorithm is to set up the objective function as a linear programming problem
- The solution of the system need not be unique.

Quantile Regression Representation

$$Q\left(p \mid X_i, \beta(p)\right) = X_i^T \beta(p)$$

$\beta(p)$ - coefficient vector, associated with p^{th}-quantile

Regression quality

- Instead of the coefficient of determination it is used its counterpart - the pseudo-R^2:

$$\hat{V}(p) = \min_{\beta(p)} \sum_i u\left(p - I(u<0)\right)\left(Y_i - \beta_0(p) - X_{i1}^T \beta(p)\right)$$

$$\bar{V}(p) = \min_{\beta(p)} \sum_i u\left(p - I(u<0)\right)\left(Y_i - \beta_0(p)\right)$$

$$R^1(p) = 1 - \frac{\hat{V}(p)}{\bar{V}(p)}$$

- Pseudo-R^2 is located between 0 and 1 and measures the regression quality for p^{th} quantile.

Quantile Regression Properties

- Robust to outliers. As long as the sign of the residual does not change, any Y_i may be changed **without shifting** the conditional quantile line.
- The regression quantiles **are correlated.**

PROPERTIES OF THE ESTIMATOR

Properties of the Estimator – 1

- Asymptotic Distribution

$$\sqrt{n}\left(\hat{\beta}_\theta - \beta_\theta\right) \xrightarrow{L} N\left(0, \Lambda_\theta\right)$$

where

$$\Lambda_\theta = \theta(1-\theta)\left(E\left[f(0 \mid x_i)x_i x_i^T\right]\right)^{-1} E\left[x_i x_i^T\right]\left(E\left[f(0 \mid x_i)x_i x_i^T\right]\right)^{-1}$$

- The covariance depends on the unknown f(.) and the value of the vector x at which the covariance is being evaluated.

Properties of the Estimator - 2

- When the error is independent of x then the coefficient covariance reduces to

$$\Lambda_\theta = \frac{\theta\left(1-\theta\right)}{f_u^{\,2}\left(0\right)}\left(E\left(xx^T\right)\right)^{-1}$$

where

$$\hat{E}(x\,x^T) = \frac{1}{n}\sum_{i=1}^{n} x_i\, x_i^T$$

Properties of the Estimator – 3

- In general the quantile regression estimator ***is more efficient than OLS***
- The efficient estimator requires knowledge of the ***true error distribution***.

Coefficient Interpretation

$$\frac{\partial Q_\theta(y_i \mid x_i)}{\partial x_{ij}}$$

The marginal change in the p^{th} conditional quantile due to a marginal change in the j^{th} element of x. There is no guarantee that the i^{th} person will remain in the same quantile after her x is changed.

Quantile Regression Hypothesis Testing

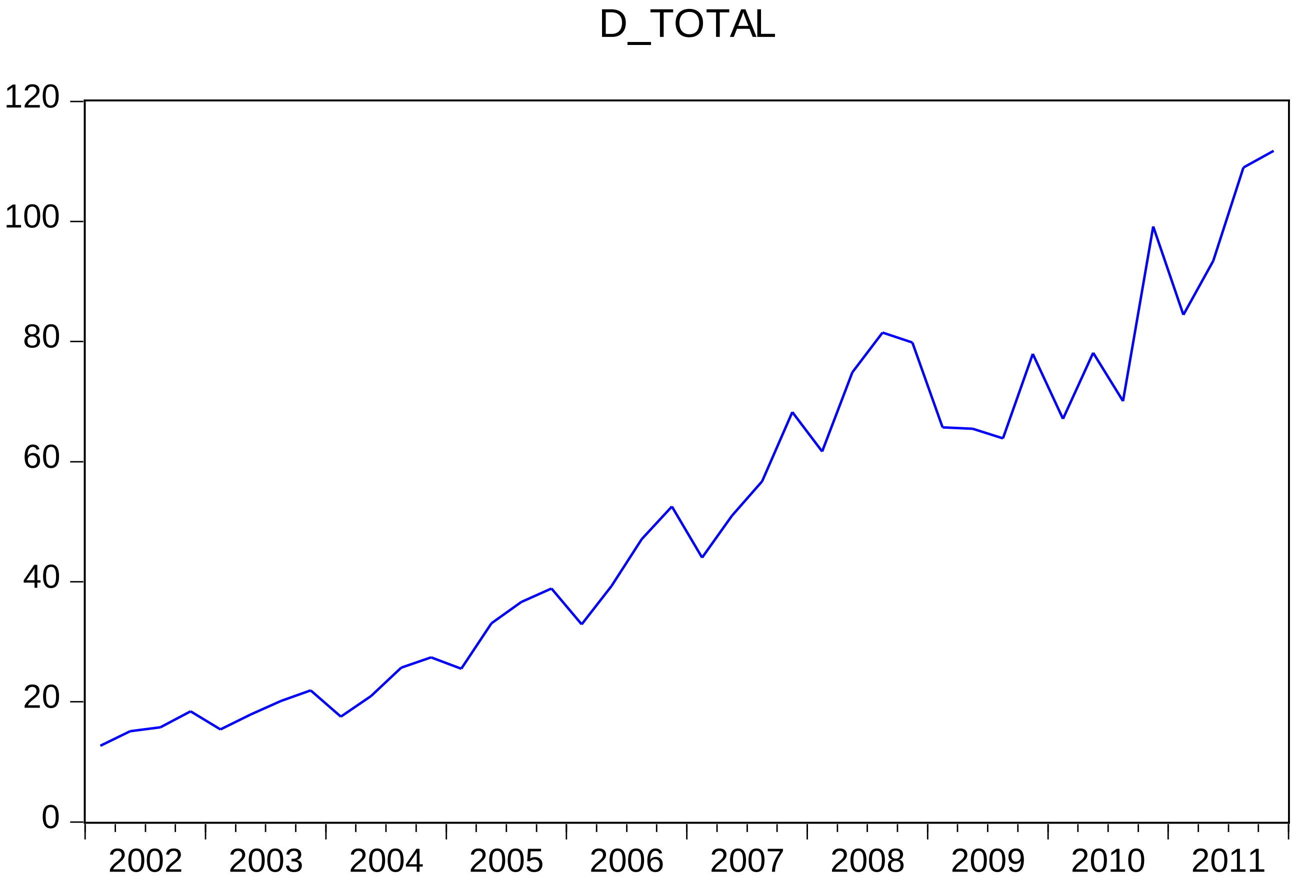
- Given asymptotic normality, one can construct asymptotic t-statistics for the coefficients
- The error term may be heteroscedastic. The test statistic is, in construction, similar to the Wald Test.
- A test for symmetry, also resembling a Wald Test, can be built relying on the invariance properties referred to above.

Heteroscedasticity

- Model: $y_i = \beta_0+\beta_1 x_i+u_i$, with iid errors.
  - The quantiles are a vertical shift of one another.
- Model: $y_i = \beta_0+\beta_1 x_i+\sigma(x_i)u_i$, errors are now heteroscedastic.
  - The quantiles now exhibit a location shift as well as a scale shift.
- Khmaladze-Koenker Test Statistic

EXAMPLE

Graph


D_TOTAL
120
100
80
60
40
20
0
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011

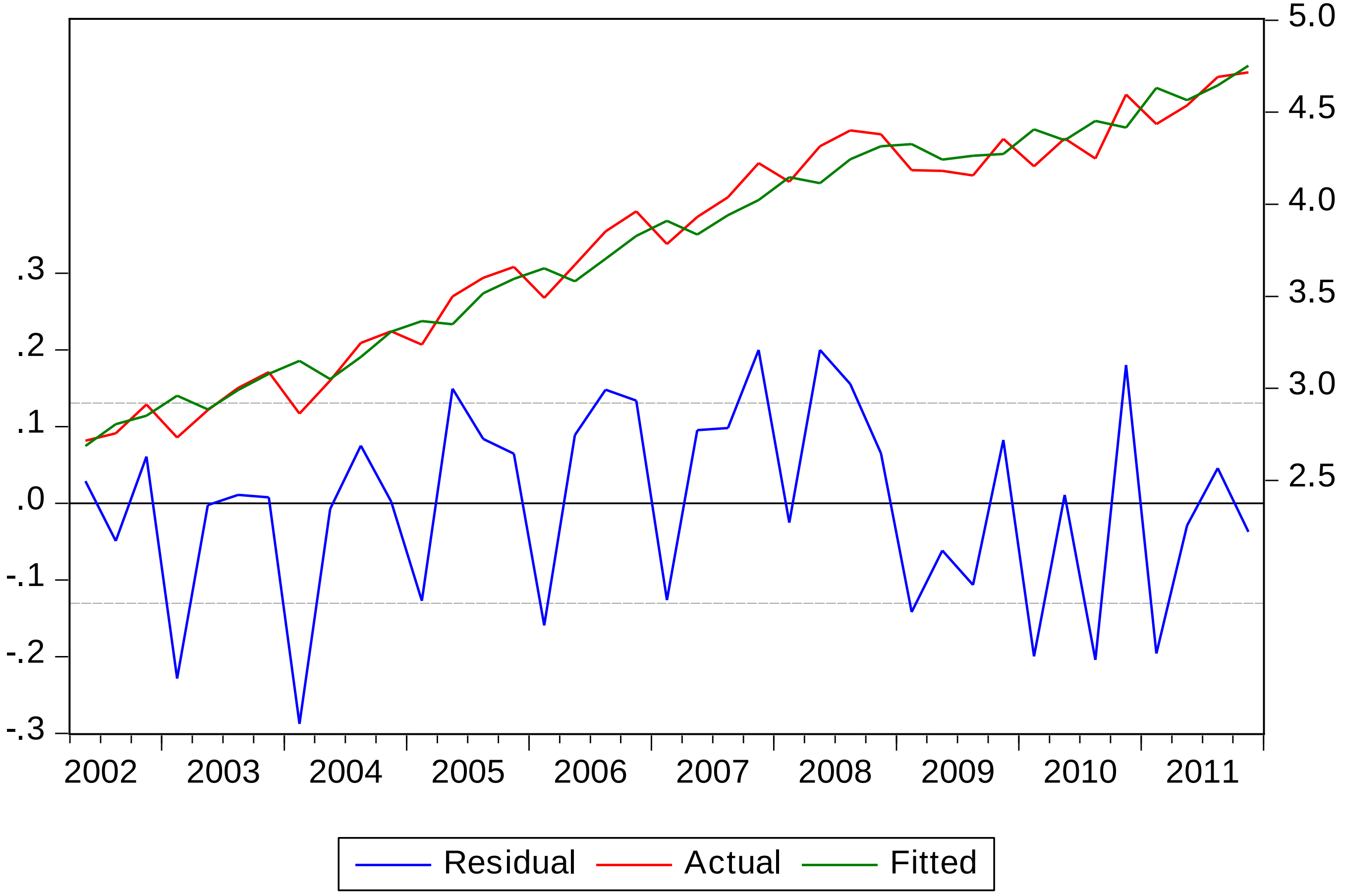
Regression Estimation (OLS)

Dependent Variable: LOG(D_TOTAL)
Method: Least Squares
Date: 12/09/12 Time: 19:33
Sample (adjusted): 2002Q2 2011Q4
Included observations: 39 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	1.259307	0.363393	3.465414	0.0014
@TREND	0.023130	0.007576	3.052982	0.0042
LOG(D_TOTAL(-1))	0.552532	0.136375	4.051565	0.0003

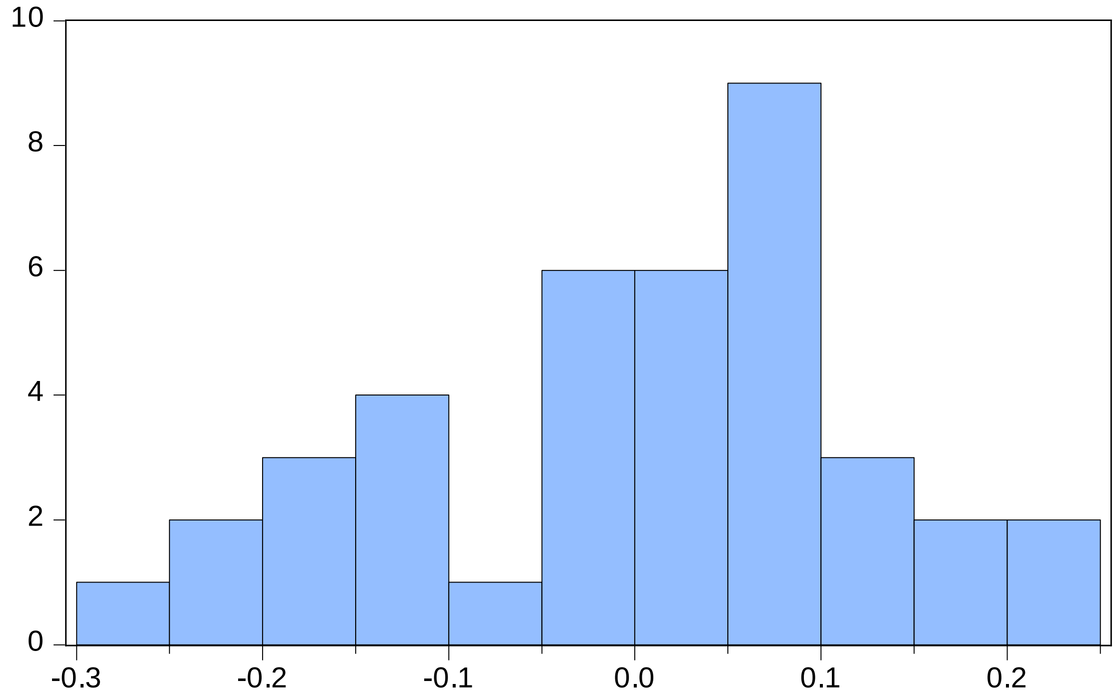
R-squared	0.958307	Mean dependent var	3.779234
Adjusted R-squared	0.955991	S.D. dependent var	0.622207
S.E. of regression	0.130529	Akaike info criterion	-1.160639
Sum squared resid	0.613362	Schwarz criterion	-1.032672
Log likelihood	25.63246	Hannan-Quinn criter.	-1.114726
F-statistic	413.7272	Durbin-Watson stat	2.235377
Prob(F-statistic)	0.000000		

Residuals



5.0
4.5
4.0
3.5
3.0
2.5
.3
.2
.1
.0
-.1
-.2
-.3
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
Residual
Actual
Fitted

Normal distribution test



Series: Residuals
Sample 2002Q2 2011Q4
Observations 39

Mean	4.86e-16
Median	0.010876
Maximum	0.200128
Minimum	-0.287605
Std. Dev.	0.127048
Skewness	-0.425670
Kurtosis	2.347786
Jarque-Bera	1.869015
Probability	0.392779

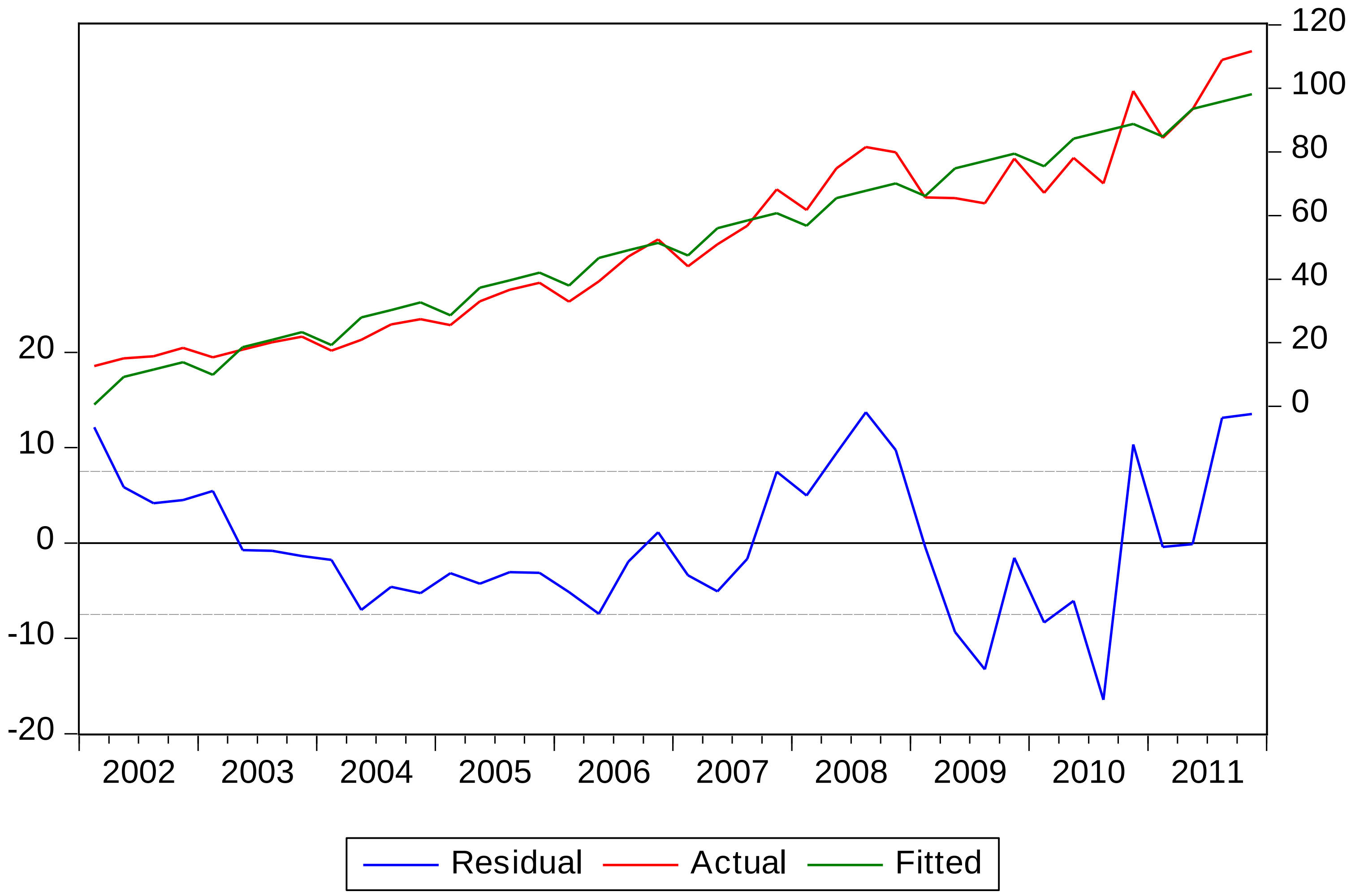
Regression Estimation (OLS)

Dependent Variable: D_TOTAL
Method: Least Squares
Date: 12/09/12 Time: 19:28
Sample: 2002Q1 2011Q4
Included observations: 40

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	6.890478	2.475560	2.783401	0.0084
@TREND	2.341282	0.103089	22.71133	0.0000
@SEAS(1)	-6.341187	2.748173	-2.307419	0.0267

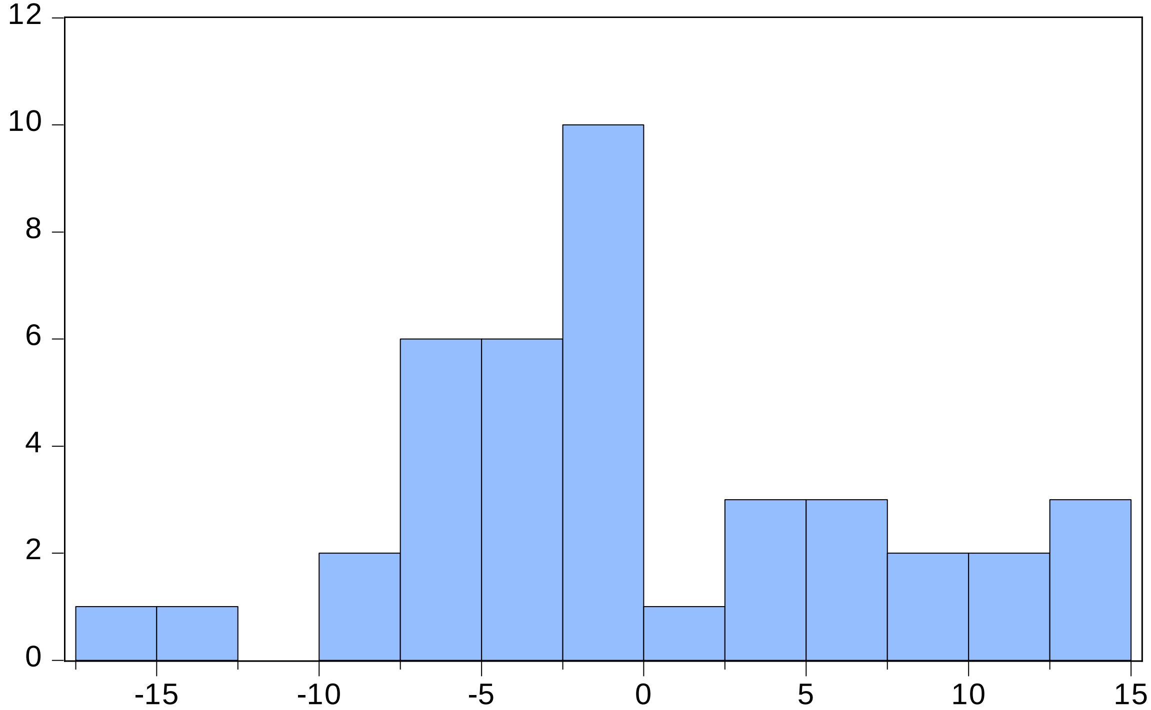
R-squared	0.934972	Mean dependent var	50.96017
Adjusted R-squared	0.931457	S.D. dependent var	28.66603
S.E. of regression	7.504971	Akaike info criterion	6.941047
Sum squared resid	2084.010	Schwarz criterion	7.067713
Log likelihood	-135.8209	Hannan-Quinn criter.	6.986845
F-statistic	265.9931	Durbin-Watson stat	0.889234
Prob(F-statistic)	0.000000		

Residuals



120
100
80
60
40
20
0
20
10
0
-10
-20
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
Residual
Actual
Fitted

Normal distribution test



Series: Residuals
Sample 2002Q1 2011Q4
Observations 40

Mean	1.55e-14
Median	-1.444590
Maximum	13.72630
Minimum	-16.42888
Std. Dev.	7.310003
Skewness	0.211776
Kurtosis	2.593100
Jarque-Bera	0.574940
Probability	0.750159

Quantile Regression Estimation

Dependent Variable: D_TOTAL
Method: Quantile Regression (tau = 0.8)
Date: 12/09/12 Time: 19:37
Sample: 2002Q1 2011Q4
Included observations: 40
Huber Sandwich Standard Errors & Covariance
Sparsity method: Kernel (Epanechnikov) using residuals
Bandwidth method: Hall-Sheather, bw=0.16717
Estimation successful but solution may not be unique

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	10.72879	3.563134	3.011054	0.0047
@TREND	2.560419	0.155048	16.51375	0.0000
@SEAS(1)	-10.47433	4.175887	-2.508290	0.0166

Pseudo R-squared	0.750713	Mean dependent var	50.96017
Adjusted R-squared	0.737238	S.D. dependent var	28.66603
S.E. of regression	11.02736	Objective	82.14514
Quantile dependent var	77.93354	Restr. objective	329.5204
Sparsity	31.50867	Quasi-LR statistic	98.13779
Prob(Quasi-LR stat)	0.000000		

Forecasting errors

Period	OLS	Quantile regression, p=0,8
1Q2012	2,39%	-0,87%
2Q2012	5,15%	-1,03%
(1Q+2Q)2012	3,79%	-0,95%

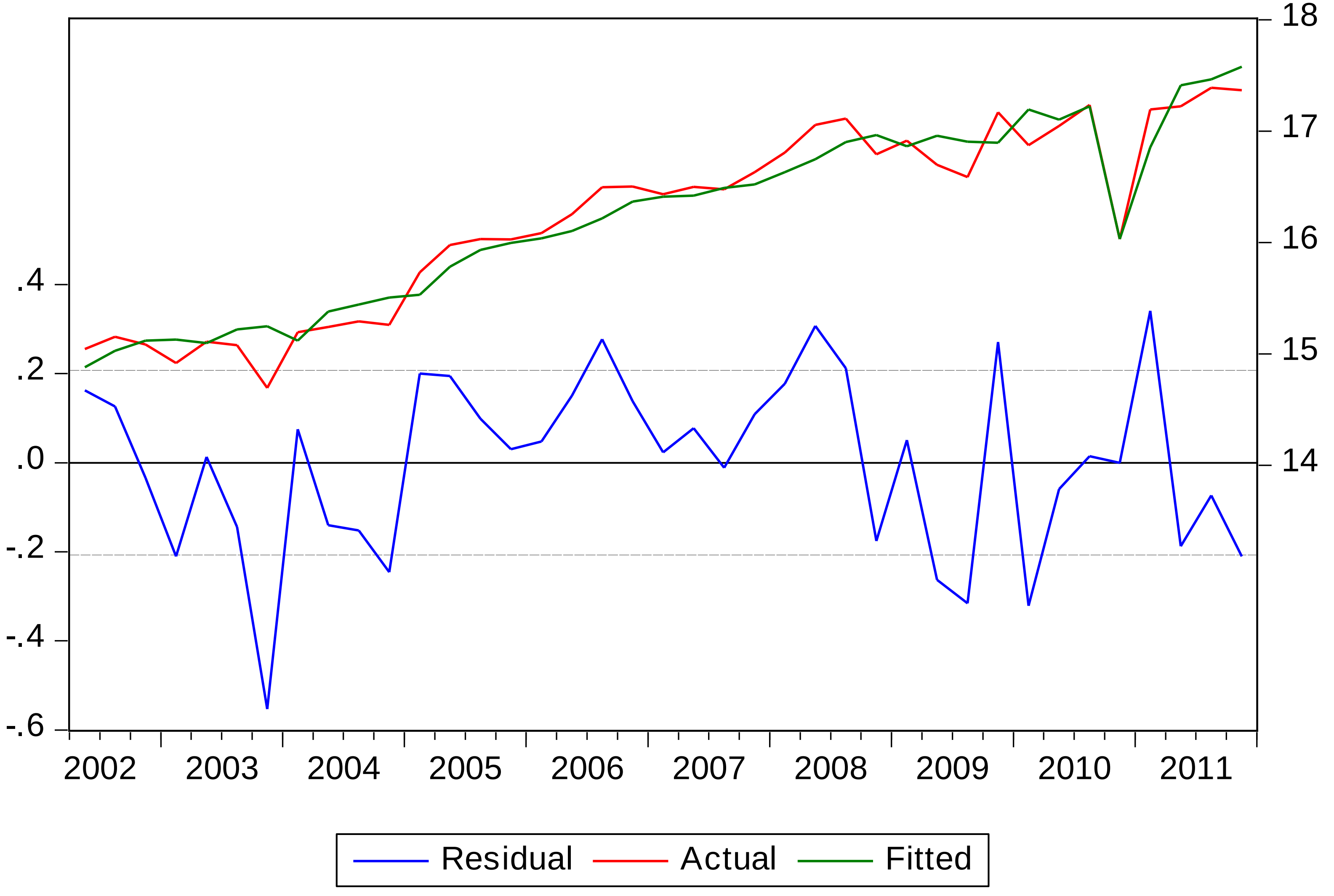
One more model

Dependent Variable: LOG(TAX_PDV)
Method: Least Squares
Date: 12/12/12 Time: 17:42
Sample (adjusted): 2002Q2 2011Q4
Included observations: 39 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	8.272613	1.584863	5.219764	0.0000
@TREND	0.040765	0.007814	5.217266	0.0000
LOG(TAX_PDV(-1))	0.443702	0.106928	4.149539	0.0002
Q	-1.315177	0.215333	-6.107633	0.0000

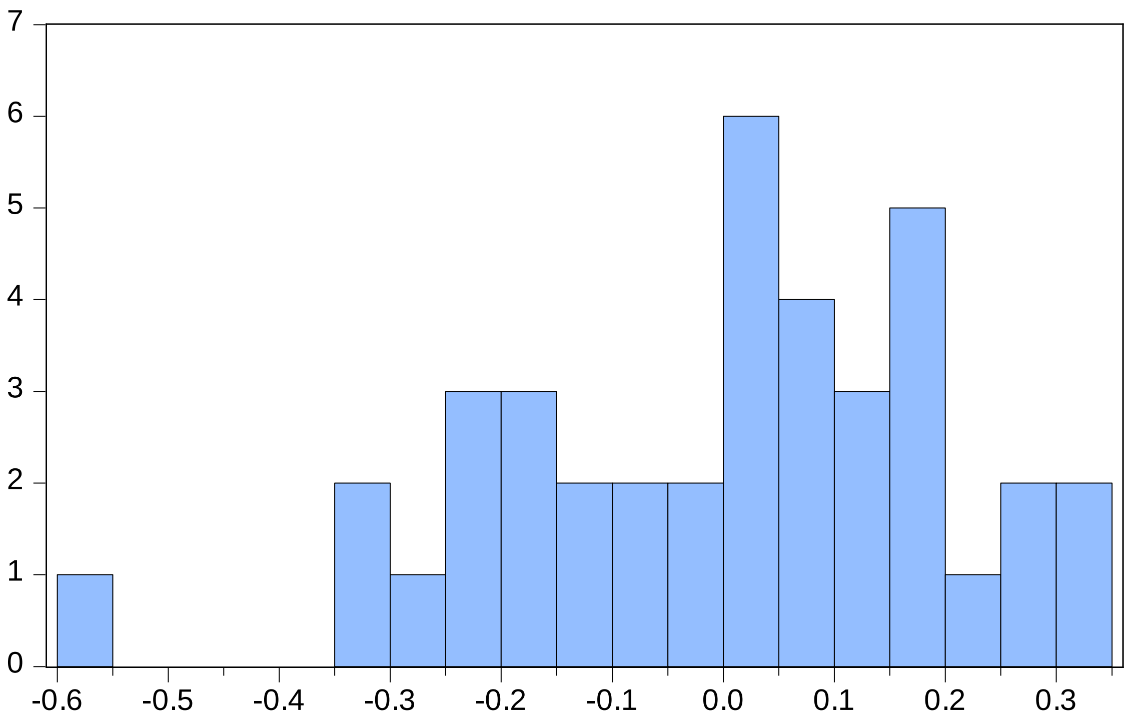
R-squared	0.941641	Mean dependent var	16.22334
Adjusted R-squared	0.936639	S.D. dependent var	0.823956
S.E. of regression	0.207403	Akaike info criterion	-0.211391
Sum squared resid	1.505561	Schwarz criterion	-0.040769
Log likelihood	8.122117	Hannan-Quinn criter.	-0.150173
F-statistic	188.2458	Durbin-Watson stat	1.711073
Prob(F-statistic)	0.000000		

Residuals



18
17
16
15
14
.4
.2
.0
-.2
-.4
-.6
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
Residual
Actual
Fitted

Normal distribution test



Series: Residuals
Sample 2002Q2 2011Q4
Observations 39

Mean	2.57e-15
Median	0.023661
Maximum	0.341068
Minimum	-0.553007
Std. Dev.	0.199048
Skewness	-0.510951
Kurtosis	2.958563
Jarque-Bera	1.699750
Probability	0.427468

Quantile regression estimation

Dependent Variable: LOG(TAX_PDV)
Method: Quantile Regression (Median)
Date: 12/09/12 Time: 20:32
Sample (adjusted): 2002Q2 2011Q4
Included observations: 39 after adjustments
Huber Sandwich Standard Errors & Covariance
Sparsity method: Kernel (Epanechnikov) using residuals
Bandwidth method: Hall-Sheather, bw=0.28649
Estimation successfully identifies unique optimal solution

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	5.611299	4.231787	1.325988	0.1934
@TREND	0.025320	0.021386	1.183912	0.2444
LOG(TAX_PDV(-1))	0.628119	0.286227	2.194481	0.0349
Q	-1.291909	0.238601	-5.414524	0.0000

Pseudo R-squared	0.777288	Mean dependent var	16.22334
Adjusted R-squared	0.758199	S.D. dependent var	0.823956
S.E. of regression	0.219122	Objective	3.001276
Quantile dependent var	16.47857	Restr. objective	13.47605
Sparsity	0.611171	Quasi-LR statistic	137.1108
Prob(Quasi-LR stat)	0.000000		

Forecasting errors

Period	OLS	Quantile regression, p=0,5
1Q2012	9,67%	-13,82%
2Q2012	28,57%	6,20%
(1Q+2Q)2012	19,02%	-3,92%

REVIEW

Problems – 1

- The distribution of Y, the "dependent" variable, conditional on the covariate X, ***may have thick tails***.
- The conditional distribution of Y may ***be asymmetric.***
- The conditional distribution of Y may ***not be unimodal***.

Problems – 2

- ANOVA and regression provide information only about the conditional mean.
- Neither regression nor ANOVA will give us robust results. Outliers are problematic, the mean is pulled toward the skewed tail, multiple modes will not be revealed.
- More knowledge about the distribution of the statistic may be important.
- The covariates may shift not only the location or scale of the distribution, they may affect the shape as well.

Reasons to use quantiles rather than means

- Analysis of distribution rather than average
- Robustness
- Skewed data
- Interested in representative value
- Interested in tails of distribution
- Unequal variation of samples

- **E.g**. Income distribution is highly skewed so median relates more to typical person that mean.

Quantile Function

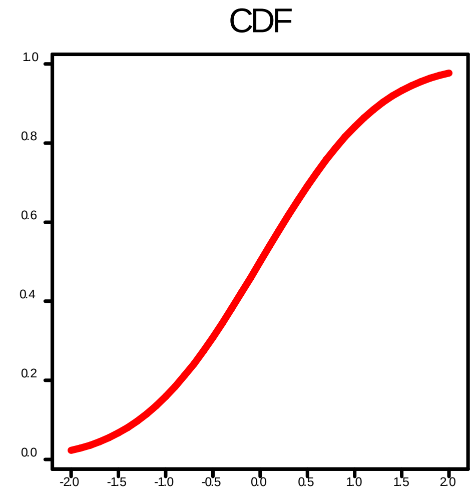
- Cumulative Distribution Function

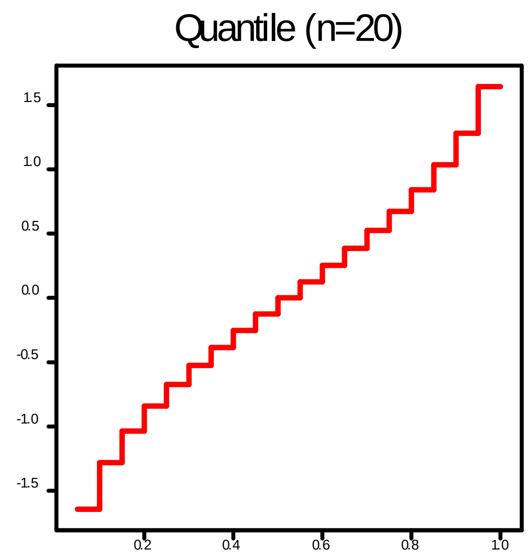
$$F(y) = \text{Prob}(Y \leq y)$$

- Quantile Function

$$Q(\tau) = \min(y : F(y) \leq \tau)$$

- Discrete step function



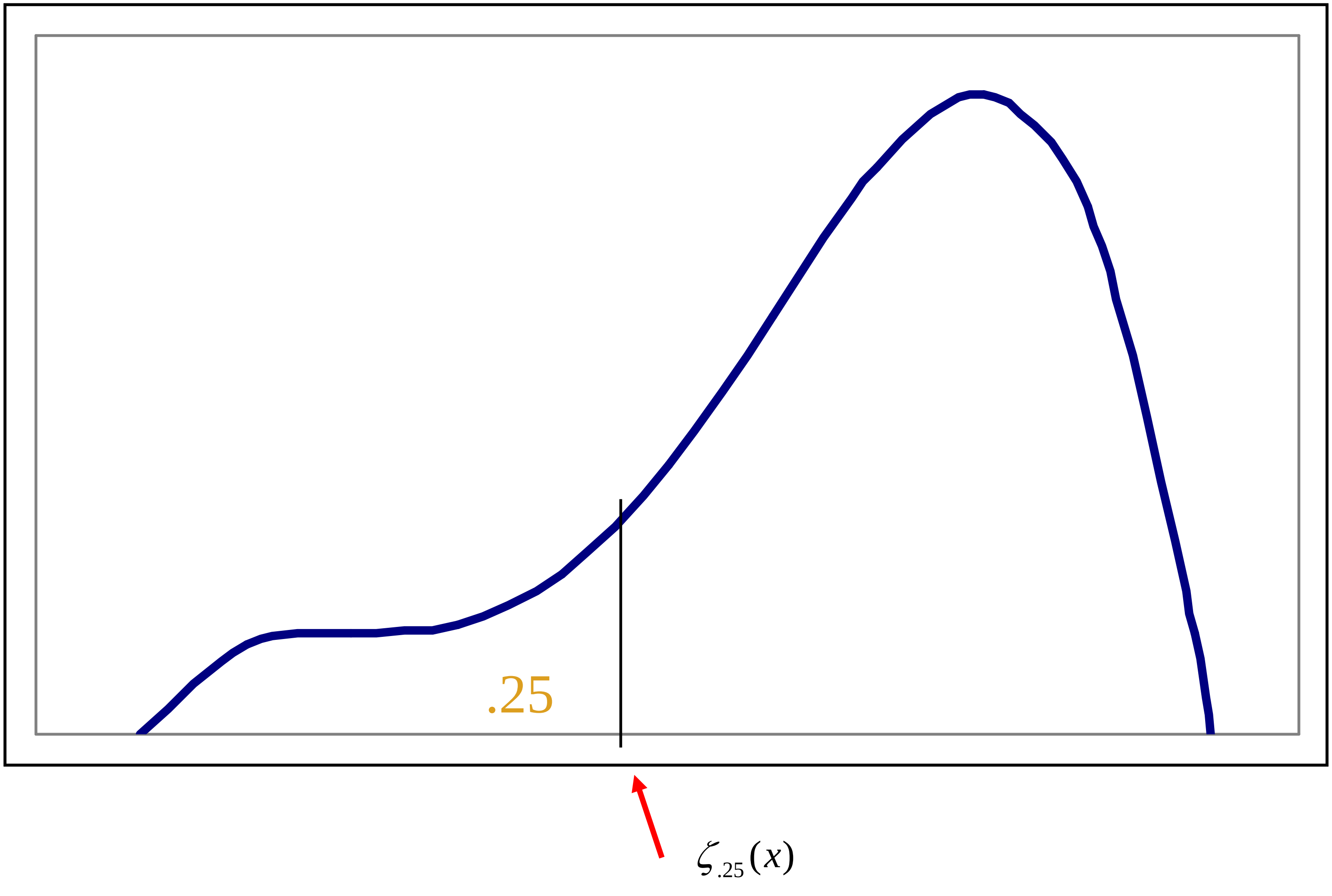


Quantile Regression Representation

$$Q\left(p \mid X_i, \beta(p)\right) = X_i^T \beta(p)$$

$\beta(p)$ - coefficient vector, associated with pth-quantile

Quantile Regression Graph



Quantile Regression Estimation

- The quantile regression coefficients are the solution to

$$\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} \left[p - \tfrac{1}{2} - \tfrac{1}{2} \operatorname{sgn}\left(y_i - x_i^T \beta \right) \right] \left(y_i - x_i^T \beta \right)$$

$$\min_{\beta} \left[\sum_{\{i | y_i \geq X_i \beta\}} p \left| y_i - X_i \beta \right| + \sum_{\{i | y_i < X_i \beta\}} (1 - p) \left| y_i - X_i \beta \right| \right]$$

Negative residuals   Positive residuals

Regression quality

- Instead of the coefficient of determination it is used its counterpart - the pseudo-R^2:

$$\hat{V}(p) = \min_{\beta(p)} \sum_i u\left(p - I(u<0)\right)\left(Y_i - \beta_0(p) - X_{i1}^T \beta(p)\right)$$

$$\bar{V}(p) = \min_{\beta(p)} \sum_i u\left(p - I(u<0)\right)\left(Y_i - \beta_0(p)\right)$$

$$R^1(p) = 1 - \frac{\hat{V}(p)}{\bar{V}(p)}$$

- Pseudo-R^2 is located between 0 and 1 and measures the regression quality for pth quantile.

Quantile Regression Properties

- Robust to outliers. As long as the sign of the residual does not change, any Y_i may be changed without shifting the conditional quantile line.
- The regression quantiles are correlated.

Coefficient Interpretation

$$\frac{\partial Q_\theta(y_i \mid x_i)}{\partial x_{ij}}$$

The marginal change in the Θ^{th} conditional quantile due to a marginal change in the j^{th} element of x. There is no guarantee that the i^{th} person will remain in the same quantile after her x is changed.

QUESTIONS?

THANK YOU FOR YOUR ATTENTION!