Lecture 10
Robust and Quantile Regression

Outliers

• Many definitions: Atypical observations, extreme values, conditional unusual values, observations outside the expected relation, etc.

• In general, we call an outlier an observation that is numerically different from the data. But, is this observation a “mistake,” say a result of measurement error, or part of the (heavy-tailed) distribution?

• In the case of normally distributed data, roughly 1 in 370 data points will deviate from the mean by 3xSD. Suppose $T=1000$. Then, 9 data points deviating from the mean by more than 3xSD indicates outliers. But, which of the 9 observations can be classified as an outliers?

• Problem with outliers: They can affect estimates. For example, with small data sets, one big outlier can seriously affect OLS estimates.
Outliers

• Several identifications methods:
  - *Eyeball*: Look at the observations away from a scatter plot.
  - *Standardized residual*: Check for errors that are two or more standard deviations away from the expected value.
  - *Leverage statistics*: It measures the difference of an independent data point from its mean. High leverage observations can be potential outliers. Leverage is measured by the diagonal values of the $P$ matrix:
    \[ h_i = \frac{1}{T} + \frac{(x_i - \bar{x})}{(T-1)s^2_x}. \]
    But, an observation can have high leverage, but no influence.
  - *Influence statistics*: Dif beta. It measures how much an observation influences a parameter estimate, say $b_j$. Dif beta is calculated by removing an observation, say $i$, recalculating $b_j$, say $b_j(-i)$, taking the difference in betas and standardizing it. Then,
    \[ Dif \ beta_j(-i) = \frac{[b_j - b_j(-i)]}{SE[b_j]}. \]

Outliers

• Deleting the observation in the upper right corner has a clear effect on the regression line. This observation has *leverage* and *influence*. 
Outliers

- A related popular influence statistic is Distance $D$ (as in Cook’s $D$). It measures the effect of deleting an observation on the fitted values, say $\hat{y}_j$.

$$D_j = \sum [\hat{y}_j - \hat{y}_j(-i)]/[K \text{MSE}]$$

where $K$ is the number of parameters in the model and MSE is mean square error of the regression model.

- The influence statistics are usually compare to some ad-hoc cut-off values used for identifying highly influential points, say $D_i > 4/T$.

- The analysis can also be carried out for groups of observations. In this case, we would be looking for blocks of highly influential observations.

Outliers: Summary of Rules of Thumb

- General rules of thumb used to identify outliers:

<table>
<thead>
<tr>
<th>Measure</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>abs(stand resid)</td>
<td>$&gt; 2$</td>
</tr>
<tr>
<td>abs(Dif Beta)</td>
<td>$&gt; 2/\sqrt{T}$</td>
</tr>
<tr>
<td>Cook's D</td>
<td>$&gt; 4/T$</td>
</tr>
<tr>
<td>leverage</td>
<td>$(2k+2)/T$</td>
</tr>
</tbody>
</table>
Outliers: Regression – SAS - Application

```sas
proc reg data = ab;
model S51-RF = MKT-RF SMB HML / white vif collinoint;
output out=capmres(keep=year S51-RF MKT-RF SMB HML r sr lev cd dffit)
   r=res student=sr h=lev cookd=cd;
run;
```

The REG Procedure
Model: MODEL1
Dependent Variable: S51-RF

<table>
<thead>
<tr>
<th>Variable</th>
<th>Label</th>
<th>DF</th>
<th>Parameter Estimate</th>
<th>Standard Error</th>
<th>t Value</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>Intercept</td>
<td>1</td>
<td>-0.33505</td>
<td>0.10816</td>
<td>-3.10</td>
<td>0.09035</td>
</tr>
<tr>
<td>XM</td>
<td></td>
<td>1</td>
<td>1.03766</td>
<td>0.02128</td>
<td>48.76</td>
<td>0.03982</td>
</tr>
<tr>
<td>SMB</td>
<td>SMB</td>
<td>1</td>
<td>1.51900</td>
<td>0.03441</td>
<td>44.15</td>
<td>0.09993</td>
</tr>
<tr>
<td>HML</td>
<td>HML</td>
<td>1</td>
<td>0.74036</td>
<td>0.03095</td>
<td>23.92</td>
<td>0.08977</td>
</tr>
</tbody>
</table>

Outliers: Distribution – SAS - Application

The UNIVARIATE Procedure
Variable: r (Studentized Residual without Current Obs)

<table>
<thead>
<tr>
<th>Location</th>
<th>Variability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>0.00097</td>
</tr>
<tr>
<td>Median</td>
<td>-0.09766</td>
</tr>
<tr>
<td>Mode</td>
<td></td>
</tr>
<tr>
<td>Interquartile Range</td>
<td>1.16010</td>
</tr>
<tr>
<td>100% Max</td>
<td>4.7696676</td>
</tr>
<tr>
<td>95%</td>
<td>1.7128005</td>
</tr>
<tr>
<td>90%</td>
<td>1.2168926</td>
</tr>
<tr>
<td>75% Q3</td>
<td>0.5638215</td>
</tr>
<tr>
<td>50% Median</td>
<td>-0.0976612</td>
</tr>
<tr>
<td>25% Q1</td>
<td>-0.5962799</td>
</tr>
<tr>
<td>10%</td>
<td>-1.1582571</td>
</tr>
<tr>
<td>5%</td>
<td>-1.4562294</td>
</tr>
<tr>
<td>0% Min</td>
<td>-3.5782300</td>
</tr>
</tbody>
</table>
### Outliers: Distribution – SAS - Application

<table>
<thead>
<tr>
<th>Histogram</th>
<th>#</th>
<th>Boxplot</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.75++</td>
<td>2</td>
<td>*</td>
</tr>
<tr>
<td>*</td>
<td>2</td>
<td>*</td>
</tr>
<tr>
<td>.*</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>.*</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>**</td>
<td>8</td>
<td>0</td>
</tr>
<tr>
<td>****</td>
<td>22</td>
<td>0</td>
</tr>
<tr>
<td>*****</td>
<td>29</td>
<td></td>
</tr>
<tr>
<td>*********</td>
<td>73</td>
<td></td>
</tr>
<tr>
<td>*************</td>
<td>132</td>
<td>+-----</td>
</tr>
<tr>
<td>****************</td>
<td>204</td>
<td>+</td>
</tr>
<tr>
<td>******************</td>
<td>247</td>
<td><em>---</em></td>
</tr>
<tr>
<td>*********************</td>
<td>168</td>
<td>+-----</td>
</tr>
<tr>
<td>**********</td>
<td>87</td>
<td></td>
</tr>
<tr>
<td>*****</td>
<td>29</td>
<td></td>
</tr>
<tr>
<td>***</td>
<td>13</td>
<td>0</td>
</tr>
<tr>
<td>.*</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>*</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>-3.75*</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* may represent up to 6 counts

### Outliers: Rules of Thumb – SAS - Application

- The histogram, Boxplot, and quantiles helps us see some potential outliers, but we cannot see which observations are potential outliers. For these, we can use Cook’s D, Diffbeta’s, standardized residuals and leverage statistics, which are estimated for each $i$.

<table>
<thead>
<tr>
<th>Observation Type</th>
<th>Proportion</th>
<th>Cutoff</th>
</tr>
</thead>
<tbody>
<tr>
<td>Outlier</td>
<td>0.0356</td>
<td>2.0000 (abs(standardized residuals) &gt; 2)</td>
</tr>
<tr>
<td>Outlier</td>
<td>0.1474</td>
<td>2/sqrt(T) (diffit &gt; 2/sqrt(1038)=0.0621)</td>
</tr>
<tr>
<td>Outlier</td>
<td>0.0501</td>
<td>4/T (cookd &gt; 4/1038=0.00385)</td>
</tr>
<tr>
<td>Leverage</td>
<td>0.0723</td>
<td>(2k+2)/T (h=leverage &gt; .00771)</td>
</tr>
</tbody>
</table>
Outliers: Rules of Thumb – SAS - Application

• What to do?
  - Use a non-linear formulation or apply a transformation (log, square root, etc.) to the data.
  - Remove suspected observations. (Sometimes, there are theoretical reasons to remove suspect observations. Typical procedure in finance, remove public utilities or financial firms from the analysis.)
  - Winsorization of the data.
  - Use dummy variables.
  - Use LAD (quantile) regressions, which are less sensitive to outliers.
  - Weight observations by size of residuals or variance (robust estimation).

• General rule: Present results with or without outliers.
Robust Estimation

- Following Huber (1981), we will interpret robustness as insensitivity to small deviations from the assumptions the model imposes on the data.

- In particular, we are interested in distributional robustness, and the impact of skewed distributions and/or outliers on regression estimates.
  - In this context, robust refers to the shape of a distribution – i.e., when the actual distribution differs from the theoretically assumed distribution.
  - Although conceptually distinct, distributional robustness and outlier resistance are, for practical purposes, synonymous.
  - Robust can also be used to describe standard errors that are adjusted for non-constant error variance. But, we have already covered this topic.

Robust Estimation – Mean vs Median

- Intuition: Under normality, OLS has optimal properties. But, under non-normality, nonlinear estimators may be better than LS estimators.

Example: i.i.d. case
Let \( \{y_i\}_1^{T} \sim \text{iid} \ F\left(\frac{y - \mu}{\sigma}\right) \) where \( F(0) = 0.5 \)

where \( F \) is a symmetric distribution with scale parameter \( \sigma \).

- Let the order statistics be \( y_{(1)} \leq \ldots \leq y_{(T)} \)
- Sample median: \( \tilde{\mu} = y_{(T+1)/2} \)
- Laplace showed that
  \[
  \sqrt{T} (\tilde{\mu} - \mu) \to N\left(0, \frac{1}{4 f(\mu = 0)^2}\right)
  \]
Robust Estimation – Mean vs Median

• Using this result, one can show:

<table>
<thead>
<tr>
<th></th>
<th>( T \text{ var}(\text{mean}) )</th>
<th>( T \text{ var}(\hat{\mu} = \text{median}) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>1</td>
<td>1.57</td>
</tr>
<tr>
<td>Laplace</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Average</td>
<td>1.5</td>
<td>1.28</td>
</tr>
</tbody>
</table>

• Intuitively, this occurs because Laplace is fat-tailed, and the median is much less sensitive to the information in the tails than the mean.

• The mean gives \( 1/T \) weight to all observations (close to the mean or in the tails). A large observation can seriously affect (influence) the mean, but not the median.

Robust Estimation – Mean vs Median

• Remark: The sample mean is the MLE under the Normal distribution; while the sample median is the MLE under the Laplace distribution.

• If we do not know which distribution is more likely, following Huber, we say the median is robust (“better”). But, if the data is normal, the median is not efficient (57% less efficient than mean).

• There are many types of robust estimators. Although they work in different ways, they all give less weight to observations that would otherwise influence the estimator.

• Ideally, we would like to design a weighting scheme that delivers a robust estimator with good properties (efficiency) under normality.
Robust Estimation – Mean vs Median

Examples: Robust estimators for central location parameter.
- The sample median, $\bar{\mu}$.
- Trimmed-Mean, the mean of the sample after fraction $\alpha$ of the largest and smallest observations have been removed.
- The “Winsorized Mean:"

$$\hat{\mu}^W = \frac{1}{T}\left((g+1)y_{(g+1)} + y_{(g+2)} + \ldots + y_{(T-g-1)} + (g+1)y_{(T-g)}\right)$$

which is similar to the trimmed-mean, but instead of throwing out the extremes, we “accumulate” them at the truncation point.
- Q: All robust, which one is better? Trade-off: robustness-efficiency.

• The concept of robust estimation can be easily extended to the problem of estimating parameters in the regression framework.

Robust Regression

• There are many types of robust regression models. Although they work in different ways, they all give less weight to observations that would otherwise influence the regression line.

• Early methods:
  - Least Absolute Deviation/Values (LAD/LAV) regression or least absolute deviation regression –i.e., minimizes $|e|$ instead of $e^2$.

• Modern methods:
  - M-Estimation
    - Huber estimates, Bi-square estimators
  - Bounded Influence Regression
    - Least Median of Squares, Least-Trimmed Squares
Review: M-Estimation

- An extremum estimator is one obtained as the optimizer of a criterion function, \( q(\mathbf{z}, \mathbf{b}) \).

Examples:
- OLS: \( \mathbf{b} = \text{arg max} (-\mathbf{e}'\mathbf{e}/T) \)
- MLE: \( \mathbf{b}_{\text{MLE}} = \text{arg max} \ln L = \sum_{i=1,...,T} \ln f(y_i, \mathbf{x}_i, \mathbf{b}) \)

- M-estimators: The objective function is a sample average or a sum.
  - "M" stands for a maximum or minimum estimators --Huber (1967).
  - It can be viewed a generalization of MLE.

- We want to obtain: \( \mathbf{b}_M = \text{argmin} \sum_i q(\mathbf{z}_i, \mathbf{b}) \) (or divided by \( T \)).
  - If \( q(y_i - \mathbf{x}_i' \mathbf{b}_M) \), \( q(.) \) measures the contribution of each residual to the objective function.

Review: M-Estimation

- We want to obtain: \( \mathbf{b}_M = \text{argmin} \sum_i q(\mathbf{z}_i, \mathbf{b}_M) \) (or divided by \( T \)).
  - In general, we solve the f.o.c. Let \( \psi = \partial q(.) / \partial \mathbf{b}' \). Then,
    \[ \sum_i \psi(y_i - \mathbf{x}_i' \mathbf{b}_M) \mathbf{x}_i = 0 \] (K equations)

  - We replace \( \psi(.) \) with the weight function, \( w_i = \psi(e_i)/e_i \)
    \[ \sum_i w_i (y_i - \mathbf{x}_i' \mathbf{b}) \mathbf{x}_i = \sum_i w_i e_i x_i = 0 \]

  These f.o.c.’s are equivalent to a weighted LS problem, which minimizes \( \sum w_i e_i^2 \).

- Q: Which \( q(.) \), or equivalently, \( w_i \) should we use to produce a robust estimator?
M-Estimation: Asymptotic Normality

• Summary
  - $\mathbf{b}_M \xrightarrow{p} \mathbf{b}_0$
  - $\mathbf{b}_M \xrightarrow{d} \mathcal{N}(\mathbf{b}_0, \text{Var}[\mathbf{b}_0])$
  - $\text{Var}[\mathbf{b}_M] = (\frac{1}{T}) \mathbf{H}_0^{-1} \mathbf{V}_0 \mathbf{H}_0^{-1}$
  - If the model is correctly specified: $\mathbf{H} = \mathbf{V}$.
  
  Then, $\text{Var}[\mathbf{b}] = \mathbf{V}_0$

- $\mathbf{H}$ and $\mathbf{V}$ are evaluated at $\mathbf{b}_0$:
  - $\mathbf{H} = \sum_i [\partial^2 q(z_i, \mathbf{b})/\partial \mathbf{b} \partial \mathbf{b}']$
  - $\mathbf{V} = \sum_i [\partial q(z_i, \mathbf{b})/\partial \mathbf{b}][\partial q(z_i, \mathbf{b})/\partial \mathbf{b}']$

M-Estimators in the Regression Context

• Many $q(z, \mathbf{b})$ can be structured to deliver a robust estimator?

• For example, we can define the family of $L_p$-estimators:
  - $q(z; \mathbf{b}) = (1/p) | x - \beta |^p$ for $1 \leq p \leq 2$
  - $s(z; \mathbf{b}) = | x - \beta |^{p-1}$
    - $x - \beta < 0$
    - $= - | x - \beta |^{p-1}$
    - $x - \beta > 0$

• Special cases:
  - $p = 2$: We get the sample mean (LS estimator for $\beta$).
    $s(z; \mathbf{b}) = \sum_i (x_i - \mathbf{b}_M) = 0$ \Rightarrow $\mathbf{b}_M = \sum_i x_i / T$

  - $p = 1$: We get the sample median as the estimator with the least absolute deviation (LAD) for the median $\beta$. (No unique solution if $T$ is even.). Numerical (linear programming) solution needed.
M-Estimators in the Regression Context: Example

<table>
<thead>
<tr>
<th>Least absolute deviations estimator</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Residuals</td>
<td>1537.58603</td>
</tr>
<tr>
<td>Sum of squares</td>
<td>6.82594</td>
</tr>
<tr>
<td>Fit</td>
<td>.98284</td>
</tr>
<tr>
<td>R-squared</td>
<td>.98180</td>
</tr>
<tr>
<td>Sum of absolute deviations</td>
<td>189.3973484</td>
</tr>
</tbody>
</table>

| Variable | Coefficient | Standard Error | b/St.Er. | P[|Z|>z] | Mean of X |
|---------|-------------|----------------|----------|--------|-----------|
| Constant| -84.0258*** | 16.08614       | -5.223   | .0000  | 9232.86   |
| Y       | .03784***   | .00271         | 13.952   | .0000  | 9232.86   |
| PG      | -17.0990*** | 4.37160        | -3.911   | .0001  | 2.31661   |

<table>
<thead>
<tr>
<th>Ordinary least squares regression</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Residuals</td>
<td>1472.79834</td>
</tr>
<tr>
<td>Standard error of e</td>
<td>6.68059</td>
</tr>
<tr>
<td>R-squared</td>
<td>.98256</td>
</tr>
<tr>
<td>Adjusted R-squared</td>
<td>.98256</td>
</tr>
</tbody>
</table>

| Variable | Coefficient | Standard Error | t-ratio | P[|T|>t] | Mean of X |
|---------|-------------|----------------|---------|--------|-----------|
| Constant| -79.7535*** | 8.67255        | -9.196  | .0000  | 9232.86   |
| Y       | .03692***   | .00132         | 28.022  | .0000  | 9232.86   |
| PG      | -15.1224*** | 1.88034        | -8.042  | .0000  | 2.31661   |

Breakdown Point: Intuition

• There are several measures of robustness of an estimator, attempting to quantify the change. One of the most commonly used is the breakdown point.

• Let X be a random sample and T(X) be an estimator. Informally, the breakdown point of the estimator is the proportion \( \frac{m}{T} \) of observations, which can be replaced by bad observations (outliers) without forcing T(X) to leave a bounded set –i.e., become infinity.

Example: The sample mean has a breakdown point equal to 0 (one observation can drive the sample mean, regardless of the other \( T-1 \) values). The median has a breakdown point 1/2 (it can tolerate 50% bad values) and \( \alpha \%-\)trimmed mean has a breakdown point \( \alpha \% \).
Breakdown Point: Definition

• Assume a sample, \( Z \), with \( T \) observations, and let \( T \) be a regression estimator. That is, we apply \( T \) to \( Z \) we get the regression coefficients:
  \[ T(Z) = b \]

• Imagine all possible “corrupted” samples \( Z^p \) that replace any subset of observations, \( m \), in the dataset with arbitrary values - i.e., influential cases.

• The maximum bias that could arise from these substitutions is:
  \[ \text{bias}(m; T, Z) = \sup_{Z'} \| T(Z') - T(Z) \| \]

• If the \( \text{bias}(m; T, Z) \) is infinite, the \( m \) outliers have an arbitrarily large effect on \( T \). In other words, the estimator breaks down.

Breakdown Point: Definition

• Then, the breakdown point for an estimator \( T \) for a finite sample \( Z \) is:
  \[ \epsilon_n^*(T, Z) = \min \left\{ \frac{m}{n}; \text{bias}(m; T, Z) \text{ is infinite} \right\} \]

• The breakdown point of an estimator is the smallest fraction of “bad” data (outliers or data grouped at the extreme of a tail) the estimator can tolerate without taking on values arbitrarily far from \( T(Z) \).

• For OLS regression one unusual case is enough to influence the coefficient estimates. Its breakdown point is then
  \[ \epsilon_n^*(T, Z) = 1/T \]

• As \( T \) gets larger, \( 1/T \) tends towards 0, meaning that the breakdown point for OLS is 0%.
Robust Regression: Methods

- Robust regression methods attempt to limit the impact of unusual cases on the regression estimates
  
  - **Least Absolute Values (LAV/LAD) regression** is robust to outliers (unusual Y values given X), but typically fares even worse than OLS for cases with high leverage.
    - If a leverage point is very far away, the LAD line will pass through it. In other words, its breakdown point is also $1/T$.
  
  - **M-Estimators** are also robust to outliers. More efficient than LAD estimators. They can have trouble handling cases with high leverage, meaning that the breakdown point is also $1/T$.
  
  - **Bounded influence methods** have a much higher breakdown point (as high as 50%) because they effectively remove a large proportion of the cases. These methods can have trouble with small samples.

Estimating the Center of a Distribution

- In order to explain how robust regression works, we start with the simple case of robust estimation of the centre of a distribution. Consider independent observations and the simple model:
  
  $$Y_i = \mu + \varepsilon_i$$

- If the underlying distribution is normal, the sample mean is the MLE.

- The mean minimizes the LS objective function:
  
  $$q_{LS} = \mathbf{e}'\mathbf{e} = \sum \varepsilon_i^2$$

- The derivative of the objective function with respect to $\varepsilon_i$ gives the influence function which determines the influence of observations:
  
  $$\psi_{LS,i}(\varepsilon) = 2 \varepsilon. \text{ That is, influence is proportional to the residual } \varepsilon_i.$$
Estimating the Center of a Distribution

- As an alternative to the mean, we consider the median as an estimator of \( \mu \). The median minimizes the LAD objective function:
  \[
  q_{LAD} = 1/T \sum_i |e_i|
  \]

- Taking the derivative of the objective function gives the shape of the influence function:
  \[
  \psi_{LAD,i}(e) =
  \begin{cases} 
  1 & \text{for } e_i > 0. \\
  0 & \text{for } e_i = 0. \\
  -1 & \text{for } e_i < 0.
  \end{cases}
  \]

- Note that influence of \( e_i \) is bounded. The fact that the median is more resistant than the mean to outliers is a favorable characteristic.

Influence Function for Mean and Median

![Figure 14.10 from Fox (1997)](image)
M-Estimation: Huber Estimates

• But, the median is far less efficient, however. If $Y \sim N(\mu, \sigma^2)$,
  \[
  \text{Var}[\bar{\mu}] = \sigma^2 / T \\
  \text{Var}[\tilde{\mu}] = \pi \sigma^2 / 2T \\
  \Rightarrow \text{The Var}[\tilde{\mu}] is \pi/2 (\approx 1.57) times as large as Var[\text{mean}].
  \]

• A good compromise between the efficiency of LS and the robustness of LAD is the Huber (1964) objective function:
  \[
  q_{\text{H,}i}(e_i) = \frac{1}{2} e_i^2 \quad \text{for } |e_i| \leq k. \quad (k = \text{tuning constant})
  \]
  \[
  = k |e_i| - \frac{1}{2} k^2 \quad \text{for } |e_i| > k.
  \]
  with an influence function:
  \[
  \psi_{\text{H,}i}(e_i) = k \quad \text{for } e_i > k.
  \]
  \[
  = e_i \quad \text{for } |e_i| \leq k.
  \]
  \[
  = -k \quad \text{for } e_i < -k.
  \]

M-Estimation: Tuning constant, $k$

• $k$ is called the tuning constant.

Note: For $k \to \infty$, the M-estimator turns into mean, for $k \to 0$, it becomes the median.

• Assuming the $\sigma=1$, setting $k=1.345$ produces 95% efficiency relative to the sample mean when the population is normal and gives substantial resistance to outliers when it is not.

• In general, $k$ is expressed as a multiple of the scale of $Y$ (the spread), $S$
  \[
  \Rightarrow k = \epsilon S.
  \]
  – We could use $\sigma$ as a measure of scale, but it is more influenced by extreme observations than is the mean.
  – Instead, we use the median absolute deviation:
  \[
  \text{MAD} = \text{median} |Y_i - \tilde{\mu}| = \text{median} |e_i|
  \]
M-Estimation: Tuning constant, $k$

- We use the median absolute deviation:
  $\text{MAD} = \text{median} |Y_i - \mu| = \text{median} |e_i|$

- The median of $Y$ serves as an initial estimate of $\mu$, thus allowing us to define $S=\text{MAD}/.6745$, which ensures that $S$ estimates $\sigma$ when the population is normal—i.e., for the standard normal $E[\text{MAD}] = 0.6745$

- Using $k=1.345$ $S$ (1.345/.6745 is about 2) produces 95% efficiency relative to the sample mean when the population is normal and gives substantial resistance to outliers when it is not.

Note: A smaller $k$ gives more resistance to outliers.

---

M-Estimation: Bi-weight Estimates

- Tukey’s bi-weight (bisquare) estimates behave somewhat differently than Huber weights, but are calculated in a similar manner

- The biweight objective function is especially resistant to observations on the extreme tails:
  \[
  q_{BW,i}(e_i) = \begin{cases} 
  k^2/6 \{1-[1-(e_i/k)^2]^3\} & \text{for } |e_i| \leq k. \\
  k^2/6 & \text{for } |e_i| > k.
  \end{cases}
  \]
  with an influence function:
  \[
  s_{BW,i}(e_i) = \begin{cases} 
  e_i \{1-[1-(e_i/k)^2]^3\} & \text{for } |e_i| \leq k. \\
  0 & \text{for } |e_i| > k.
  \end{cases}
  \]

- For this function, $k=4.685$ $S$ (4.685/.6745 about 7 MADS) produces 95% efficiency when sampling from a normal population.
M-Estimation and Regression

- Since regression is based on the mean, it is easy to extend the idea of M-estimation to regression. The linear model is:
  \[ y_i = x_i' b + \varepsilon_i \]

- The M-estimator then minimizes the objective function:
  \[ q = \sum_i q(y_i - x_i' b) \]
  with f.o.c.’s:
  \[ \sum_i \psi(y_i - x_i' b) x_i = 0 \]

- We have a system of \( K \) equations. We replace \( \psi(\cdot) \) with the weight function, \( w(\varepsilon) = \psi(\cdot)/\varepsilon \):
  \[ \sum_i w_i (y_i - x_i' b) x_i = 0 \]

- The solution assigns a different weight to each case depending on the size of their residual; similar to a weighted least squares problem.

M-Estimation and Regression: Loss Functions

- Different loss functions:
M-Estimation and Regression: Weights

- The weight function: \( w(e) = \psi(.) / e \):

<table>
<thead>
<tr>
<th>Method</th>
<th>Objective Function</th>
<th>Weight Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Least-Squares</td>
<td>( \rho_{LS}(e) = e^2 )</td>
<td>( w_{L}(e) = 1 )</td>
</tr>
<tr>
<td>Huber</td>
<td>( \rho_H(e) = \begin{cases} \frac{1}{2}e^2 &amp; \text{for }</td>
<td>e</td>
</tr>
<tr>
<td>Bisquare</td>
<td>( \rho_B(e) = \begin{cases} \frac{k^2}{e^2} \left( 1 - \left( \frac{e}{k} \right)^2 \right)^2 &amp; \text{for }</td>
<td>e</td>
</tr>
</tbody>
</table>

Weight Functions for Various Estimators

![Graphs showing weight functions for various estimators](image)
M-Estimation and Regression: Algorithm

- The solution assigns a different weight to each case depending on the size of their residual, and thus minimizes the weighted sum of squares.
  \[ \sum w_i e_i^2 = 0 \]

- The \( w_i \) weights depend on the residuals in the model. An iterative solution (using Iterative Re-weighted Least Squares, IRLS) is needed.

- The solution to this problem is weighted LS:
  1. Set initial \( b_0 \), say by using OLS. Get \( e_0 \).
  2. Estimate the scale of the residuals \( S^0 \) and the weights \( w_i^0 \).
  3. Estimate \( b_j \): 
     \[ b_j = (X'WX)^{-1} X'Wy \quad W = \text{diag}\{w_i^{j-1}\} \]
  4. With \( b_j \) go back to (1). Repeat steps (1)-(3) until convergence.

M-Estimation and Regression

- Usual weight functions: Huber and Biweight (bisquare) weights.

- M-Estimators are statistically equally efficient as OLS if the distribution is normal, while at the same time are more robust with respect to influential cases.

- However, M-estimation can still be influenced by a single very extreme X-value—i.e., like OLS, it still has a breakdown point of 0.
Bounded Influence Regression: LTS

- M-estimation can still be influenced by a single very extreme X-value—i.e., like OLS, it still has a breakdown point of 0

- *Least-trimmed-squares* (LTS) estimators – see Rousseeuw (1984) – can have a breakdown point up to 50% -i.e., half the data can be influential in the OLS sense before the LTS estimator is seriously affected.

  – Least-trimmed-squares essentially proceeds with OLS after eliminating the most extreme positive or negative residuals.

- LTS orders the squared residuals from smallest to largest: \((e^2)(1), (e^2)(2), \ldots, (e^2)(T)\)

- Then, LTS calculates \(b\) that minimizes the sum of only the smaller half of the residuals.

\[
\sum_{i=m}^{T}(e^2)_{i} = 0
\]

where \(m = [T/2] + 1\); the square bracket indicates rounding down.

- By using only the 50% of the data that fits closest to the original OLS line, LTS completely ignores extreme outliers. The breakdown value for the LTS estimate is \((T-m)/T\).

- On the other hand, this method can misrepresent the trend in the data if it is characterized by clusters of extreme cases or if the data set is relatively small.
Bounded Influence Regression: LMS

• An alternative bounded influence method is Least Median Squares (LMS).

• Rather than minimize the sum of the least squares function, this model minimizes the median of the squared residuals, $e_i^2$.

• The breakdown value for the LTS estimate is also $(T-m)/T$.

• LMS is very robust with respect to outliers both in terms of X and Y.

• But, it performs poorly from the point of view of asymptotic efficiency. Also, relative to LMS, LTS’s objective function is smoother, making the LTS estimate less jumpy -i.e., less sensitive to local effects.

Robust Regression: Application 1

De Long and Summers (1991) studied the national growth of 61 countries from 1960 to 1985 using OLS:

$$\text{GDP}_i = \beta_0 + \beta_1 \text{LFG}_i + \beta_2 \text{GAP}_i + \beta_3 \text{EQP}_i + \beta_4 \text{NEQ}_i + \varepsilon_i$$

where GDP growth per worker (GDP) and the regressors are labor force growth (LFG), relative GDP gap (GAP), equipment investment (EQP), and nonequipment investment (NEQ).

• The OLS analysis: GAP and EQP have a significant effect on GDP at the 5% level.
Robust Regression: Application 1

Zaman, Rousseeuw, and Orhan (2001) used robust techniques to estimate the same model (Zambia (observation #60) an outlier):

\[ \text{GDP}_i = \beta_0 + \beta_1 \text{LFG}_i + \beta_2 \text{GAP}_i + \beta_3 \text{EQP}_i + \beta_4 \text{NEQ}_i + \varepsilon_i \]

- Huber M-estimates: Besides GAP and EQP, the robust analysis also show NEQ has significant effect on GDP.

Robust Regression: Diagnostics

- It is common to analyze the residuals for outliers (as usual) and leverage points. To check for leverage points, Rousseeuw (1984) proposes a robust version of the Mahalanobis distance by using a generalized minimum covariance determinant (MCD) method.

- Mahalanobis Distance is the square root of a standard Wald distance:

\[ MD(x_i) = [(x_i - \bar{x})^T \hat{C}(X)^{-1} (x_i - \bar{x})]^{1/2} \]

where \( \bar{x} \) is the mean and \( \hat{C}(X) \) is the variance (scale or scatter) of \( X \).

- Rousseeuw's Robust Distance is given by

\[ RD(x_i) = [(x_i - T(X))^T C(X)^{-1} (x_i - T(X))]^{1/2} \]

where \( T(X) \) and \( C(X) \) are the robust multivariate location and scale, respectively, obtained by MCD.
Robust Regression: LTS - Application 1

Analysis of robust residuals. Lots of leverage observations, but only one outlier (Zambia, #60).

The analysis of robust residuals revealed Zambia (#60) as an outlier. Potentially, this can create problems for M-estimators. LTS estimation is has a better breakdown point.
Robust Regression: LTS - Application 1

After removing the outlier (Zambia), we re-estimate model:

![Figure 7. Final Weighted LS estimates](image)

Robust Regression: 3 Factor Model - Application 2

We run the 3 Fama-French factor model, for the lowest size (decile) portfolio.

```plaintext
proc robustreg data=ab;
model y1 = xm SMB HML;
output out=robout r=resid sr=stdres;
run;
```

Parameter Estimates

<table>
<thead>
<tr>
<th>Parameter</th>
<th>DF</th>
<th>Estimate</th>
<th>SE</th>
<th>Limits</th>
<th>Chi-Square</th>
<th>Pr &gt; ChiSq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1</td>
<td>-0.4184</td>
<td>0.0637</td>
<td>-0.5432 -0.2937</td>
<td>43.20</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>xm</td>
<td>1</td>
<td>0.9660</td>
<td>0.0125</td>
<td>0.9415 0.9906</td>
<td>5947.57</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>SMB</td>
<td>1</td>
<td>1.2760</td>
<td>0.0203</td>
<td>1.2363 1.3157</td>
<td>3969.66</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>HML</td>
<td>1</td>
<td>0.4478</td>
<td>0.0182</td>
<td>0.4121 0.4835</td>
<td>604.18</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Scale</td>
<td>1</td>
<td>1.8128</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Robust Regression: Remarks

• Separated points can have a strong influence on statistical models
  – Unusual cases can substantially influence the fit of the OLS model. Cases that are both outliers and high leverage exert influence on both the slopes and intercept of the model
  – Outliers may also indicate that our model fails to capture important characteristics of the data

• Efforts should be made to remedy the problem of unusual cases before proceeding to robust regression

• If robust regression is used, careful attention must be paid to the model—different procedures can give completely different answers.

Robust Regression: Remarks

• No one robust regression technique is best for all data

• There are some considerations, but even these do not hold up all the time:
  – LAD regression should generally be avoided because it is less efficient than other techniques and often not very resistant
  – Bounded influence regression models, which can have a breaking point as high as 50%, often work very well with large datasets. But, they tend to perform poorly with small datasets.

• M-Estimation is typically better for small datasets, but its standard errors are not reliable for small samples. This can be overcome by using bootstrapping to obtain new estimates of the standard errors.
Quantile Regression

- Mosteller and Tukey (1977):
  “What the regression curve does is a grand summary for the the averages of the distributions corresponding to the set of x’s. We could go further and compute several different regression curves corresponding to the various percentage points of the distribution and thus get a more complete picture.”

- One might be interested in behavior of say, lower tail of the conditional distribution rather than in its mean.

- For example, how does a 1% increase in market returns affect the returns of small size firms?

Quantiles: Characterizing a Distribution

- We are used to assume a distribution and describe it through its moments: mean, variance, skewness, etc. Some distributions are characterized by few parameters. For example, the normal is completely described by the mean and the variance.

- A different approach. Use quantiles instead. For example:
  - Median
  - Interquartile Range
  - Interdecile Range
  - Symmetry = $(\zeta_{.75} - \zeta_{.5})/(\zeta_{.5} - \zeta_{.25})$
  - Tail Weight = $(\zeta_{.90} - \zeta_{.10})/(\zeta_{.75} - \zeta_{.25})$
Quantiles

Quantiles
• We say that a firm is in the $\theta^k$ quantile if it is bigger than the proportion $\theta_k$ of the reference group of firms, and smaller than the proportion $(1-\theta_k)$.

• The $\theta^k$ sample quantile is simply $y(k)$, where $k$ is the smallest integer such that $K/T<\theta$. (Note the relation between rank and quantile.)

Quantiles: Definition

Definition:
(1) Discrete RV. Given $\theta \in [0, 1]$. A $\theta$th quantile of a discrete RV $Z$ is any number $\zeta_\theta$ such that $\Pr(Z<\zeta_\theta) \leq \theta \leq \Pr(Z \geq \zeta_\theta)$.

Example: Suppose $Z=\{3, 4, 7, 9, 9, 11, 17, 21\}$ and $\theta = 0.5$ then $\Pr(Z<9) = 3/8 \leq 1/2 \leq \Pr(Z \geq 9) = 5/8$.

(2) Continuous RV. Let $Z$ be a continuous r.v. with cdf $F()$, then $\Pr(Z<z) = \Pr(Z \leq z) = F(z)$ for every $z$ in the support and a $\theta$th quantile is any number $\zeta_\theta$ such that $F(\zeta_\theta) = \theta$.

• If $F$ is continuous and strictly increasing then the inverse exists and $\zeta_\theta = F^{-1}(\theta)$. 
Quantiles: CDF and Quantile Function

- Cumulative Distribution Function
  \[ F(y) = \text{Prob}(Y \leq y) \]

- Quantile Function
  \[ Q(\theta) = \min(y : F(y) \leq \theta) \]

=> Discrete step function

Quantiles

- It can be shown that quantile \((\theta)\) is the solution to
  \[ \arg \min_{\xi} \frac{1}{T} \left\{ \sum_{y_i \geq \xi} \theta |y_i - \xi| + \sum_{y_i < \xi} (1 - \theta) |y_i - \xi| \right\} \]

- If \(\theta = 1/2\), then this becomes
  \[ \arg \min_{\mu} \frac{1}{T} \sum_{i=1}^{T} |y_i - \xi| \]
  which yields a f.o.c.:
  \[ 0 = (-1/T) \sum_{i=1}^{T} \text{sgn}(y_i - \xi) \]
  where \(\text{sgn}\) ("signum") function:
  \[ \text{sgn}(\nu) = 1 - 2 I[\nu < 0] \]
  (defined to be right-continuous).

=> the sample median, \(\xi_{0.50}\), solves this problem (easier to visualize with expectations).
Quantile Regression

• Basset and Koenker (1978, JASA) suggest simply replacing the $\xi$ in the definition of the quantile estimator

$$\arg\min_{\xi} \sum_{y_i \geq \xi} \theta |y_i - \xi| + \sum_{y_i < \xi} (1 - \theta) |y_i - \xi|$$

with $X', \beta$ to get the quantile regression

$$\arg\min_{\beta} \sum_{y_i \geq X', \beta} \theta |y_i - X', \beta| + \sum_{y_i < X', \beta} (1 - \theta) |y_i - X', \beta| = \sum_{\epsilon_i \geq 0} \theta |\epsilon_i| + \sum_{\epsilon_i < 0} (1 - \theta) |\epsilon_i|$$

• If $\theta = 1/2$, then this becomes LAD estimation. We have a symmetric weighting of observations with positive and negative residuals. But, if $\theta \neq 1/2$, the weighting is asymmetric.

Quantile Regression

• We define a family of regressions:

$$\zeta_\theta = Q(y| x, \theta) = X'\beta_\theta, \quad \theta \in [0,1]$$

- Median regression is obtained by setting $\theta = .50$:

$$\zeta_{\theta=.50} = Q(y| x, .50) = X'\beta_{.50}$$
Quantile Regression

**Note:** Median regression estimated by LAD. It estimates the same parameters as OLS if symmetric conditional distribution.

- We assume correct specification of the quantile, \( Q(y | x, \theta) = X'\beta \). That is, \( X'\beta \) is a particular linear combination of the independent variables such that

\[
\theta = \Pr(Y \leq \zeta_\theta(X) | X) = \Pr(Y \leq X\beta) = F(\zeta_\theta(X) | X)
\]

Q: Why use quantile (median) regression?
- Semiparametric
- Robust to some extensions (heteroscedasticity?)
- Complete characterization of conditional distribution.
Quantile Regression: Loss Function

- Different from LS, now we minimize an asymmetric absolute loss function, given by

$$\arg \min_{\beta} \rho_{\theta}(y_t, X_t, \beta) = \arg \min_{\beta} \sum_{y_t \geq X_t \beta} \theta |y_t - X_t \beta| + \sum_{y_t < X_t \beta} (1 - \theta) |y_t - X_t \beta|$$

for some $\theta$.

- We call $\rho_{\theta}$ the tilted absolute value function. It is convex. The local minimum is a global one, which assures uniqueness (and identification).

Quantile Regression: Loss Function

Absolute Loss vs. Quadratic Loss over errors

A quadratic loss penalizes large errors very heavily. When $p=.5$ our best predictor is the median; it does not give as much weight to outliers. When $p=.7$ the loss is asymmetric; large positive errors are more heavily penalized than negative errors.
Quantile Regression: Estimation

• Optimization problem:
\[
\min_{\beta} \sum_{\varepsilon_i \geq 0} \theta |\varepsilon_i| + \sum_{\varepsilon_i < 0} (1 - \theta) |\varepsilon_i| = \sum_{i=1}^{T} (\theta - I[y_i < \theta]) \varepsilon_i
\]

• Simple intuition: number of negative residuals $\leq T \theta \leq$ number of negative residuals + number of zero residuals.

• Since the loss function is piecewise linear, solving it is a linear programming problem. Trick: replace absolute values by positivity constraints. That is,
\[
\min \{ \sum_{i=1}^{T} \theta \varepsilon_i^+ + (1 - \theta) \varepsilon_i^- = \theta t^+ \varepsilon^+ + (1 - \theta) t^- \varepsilon^- \}
\]
subject to:
\[
y = X\beta + \varepsilon^+ - \varepsilon^- \quad (\varepsilon_i^- \leq y_i - X_i \beta \leq \varepsilon_i^+)\]
\[
\varepsilon_i^+ \geq 0, \quad \varepsilon_i^- \geq 0
\]

Quantile Regression: Estimation

• The usual software packages will use the Barrodale and Roberts (1974) simplex algorithm or a Frisch-Newton (FN) algorithm.

• For large data sets, the FN method is used. It combines a log-barrier Lagrangian (Frisch part) with steepest descent steps (Newton part). For very large data sets, FN algorithm is combined with a preprocessing step, which makes the computations faster.

• Solution at vertex of feasible region. The solution need not be unique (along the edge). The fitted line will go through $k$ data points.

• Well known program in R, written by Koenker and described in Koenker’s Vignette article (2005).
Quantile Regression: Optimality

• Proposition

Under the asymmetric absolute loss function \( q_\theta \) a best predictor of \( Y \) given \( X=x \) is the \( \theta \)th conditional quantile, \( \zeta_\theta \).

**Example:** Let \( \theta = .5 \). Then, the best predictor is the median fitted value.

• That is, under asymmetric absolute loss, the quantile regression estimator is more efficient than OLS.

• We offer this without proof. The proof would be similar in construction to the Gauss-Markov Theorem, which states that the conditional mean is best linear unbiased.

Properties of the Estimator

• Consistency

Consistency of \( \hat{\beta}_\theta \) is easy. The minimand \( S_n(.) \) is continuous in \( \beta \) with probability 1. In fact, \( S_n(.) \) is convex in \( \beta \); then, consistency follows if \( S_n \) can be shown to converge pointwise to a function that is uniquely minimized at the true value \( \beta_\theta \).

• To prove consistency, we impose conditions on the model:
  1. The data \((x_i, y_i)\) are i.i.d. across \( i \)
  2. The regressors have bounded second moment.
  3. \( \varepsilon_i | X_i \) is continuously distributed; with conditional density \( f(\varepsilon_i | X_i) \) satisfying the conditional quantile restriction.
  4. The regressors and error density satisfy a local identification condition: \( C = E[\varepsilon_i(0) xx'] \) is a pd matrix.
Properties of the Estimator

• Asymptotic Normality (under i.i.d assumption)
  The lack of continuously differentiable $S_n(\beta)$ complicates the usual derivation of asymptotic normality (through Taylor’s expansion).

• But, an approximate f.o.c. can be used -through $\text{sgn}(.)$. Additional conditions (stochastic equicontinuity) need to be established before using the Lindeberg-Levy CLT, which establishes:

$$\sqrt{T}(\hat{\beta}_0 - \beta_0) \xrightarrow{d} N(0, \Lambda_0)$$

where

$$\Lambda_0 = \theta(1-\theta) \left( E[f_\varepsilon(0|x_i)x_i'x_i']^{-1} E[x_i'x_i'] \right) \left( E[f_\varepsilon(0|x_i)x_i'x_i']^{-1} E[x_i'x_i'] \right)^{-1}$$

• We have a sandwich estimator. The variance matrix depends on the unknown $f_\varepsilon(.|x)$ and the $X$, at which the covariance is being evaluated.

Properties of the Estimator

• We need to estimate $E[f_\varepsilon(0|x)xx']$, complicated without knowing $f_\varepsilon(.|x)$! It can be done through non-parametric kernel estimation.

• When the error is independent of $x$ –i.e., $f_\varepsilon(\varepsilon_i|X_i) = f_\varepsilon(\varepsilon_i)$–, then the coefficient covariance reduces to

$$\Lambda_0 = \frac{\theta (1-\theta)}{f_\varepsilon^2(0)} \left( E(xx') \right)^{-1}$$

where

$$\hat{E}(xx') = \frac{1}{n} \sum_{i=1}^n x_i x_i'$$

• The variance is related to a Bernoulli variance $[\theta(1-\theta)]$ –divided by the square density of $Y$ at the quantile, analogous to a sample size.
Properties of the Estimator

• The previous results can be extended to multivariate cases –i.e., joint estimates of several quantiles. We obtain convergence to a multivariate normal distribution.

• In general, the quantile regression estimator is more efficient than OLS. But, efficiency requires knowledge of the true error’s pdf.

• Robust to outliers. As long as the sign of the residual does not change, any \( y_i \) can be arbitrarily changed without shifting the conditional quantile line.

• The regression quantiles are correlated.

Partial Effects and Prediction

• The marginal change in the \( \Theta \)th conditional quantile due to a marginal change in the \( j \)th element of \( x \).

\[
\frac{\partial Q_\theta (y_i | X_i)}{\partial x_{i,j}}
\]

• Under linearity, the effect will be \( \beta_j \). But, if non-linearities are included, the partial effect will be a function of \( x \).

**Note:** There is no guarantee that the \( i \)th observation will remain in the same quantile after \( x_{i,j} \) changes.

• Using \( \hat{\beta}_\theta \) and \( X \) values, predicted values of \( \hat{y}_{i\theta} \) can be computed. Suppose we have \( X = x_{i\theta} \), the predicted 90\textsuperscript{th} quantile is \( x_{i\theta}'\hat{\beta}_{90} \).
Hypothesis Testing: Standard Errors

• Given asymptotic normality, one can construct asymptotic t-statistics for the coefficients. But which standard errors should be used?

• We can use the asymptotic estimator, but in non-i.i.d. situations is complicated. Inversion of a rank test --Koenker (1994, 1996)-- can be used to construct C.I.’s in a non-i.i.d. error context.

• Bootstrapping works well. Parzen, Wei, and Ying (1994) have suggested that rather than bootstrapping (x; y) pairs, instead bootstrap the quantile regression gradient condition. It produces a pivotal approach.

Hypothesis Testing

• Alternatively, confidence regions for the quantile regression parameters can be computed from the empirical distribution of the sample of bootstrapped $b_j(\theta)$’s, the so-called percentile method.

• These procedures can be extended to deal with the joint distribution of several quantile regression estimators $\{b_j(\theta_k), k = 1,2,\ldots, K\}$. This would be needed to test equality of slope parameters across quantiles.

• The error term may be heteroscedastic. Efficiency issue. There are many tests for heteroscedasticity in this context.

• A test for symmetry, resembling a Wald Test, can be constructed which could not be done under Least Squares estimation.
Crossings

- Since quantile regressions are typically estimated individually, the quantile curves can cross, leading to strange (an invalid) results.

- Crossings problems increase with the number of regressors.

- Simultaneous estimation, with constraints are one solution.

- Individual specification of each quantile also works. For example:

  \[ y = X\beta_0 + \varepsilon^0, \quad P[\varepsilon^0 < 0 \mid X] = \theta_0 \quad \text{(say, } \theta_0 = .5) \]

  \[ y = X\beta_0 - \exp(X\beta_1) + \varepsilon^1, \quad P[\varepsilon^1 < 0 \mid X] = \theta_1 \quad \text{(say, } \theta_0 = .25) \]

  \[ y = X\beta_0 + \exp(X\beta_2) + \varepsilon^2, \quad P[\varepsilon^2 < 0 \mid X] = \theta_2 \quad \text{(say, } \theta_0 = .75) \]

  Note: Since \( \exp(\cdot) \) is positive, the quantiles by design never cross.

Quantile Linear Regression: Application 1

Food Expenditure vs Income

Engel 1857 survey of 235 Belgian households

Q: Change of slope at different quantiles?
**Quantile Linear Regression: Application 1**

- **Note:** Variation of Parameter with Quantiles.

**Quantile Linear Regression: Application 2**

<table>
<thead>
<tr>
<th>Estimator</th>
<th>Coefficient</th>
<th>Standard Error</th>
<th>z</th>
<th>Prob [z &gt; Z]</th>
<th>95% Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-0.72107**</td>
<td>0.01391</td>
<td>-5.12</td>
<td>0.00000</td>
<td>-1.1163 - 0.3745</td>
</tr>
<tr>
<td>AGE</td>
<td>0.00396**</td>
<td>0.00416</td>
<td>0.96</td>
<td>0.33333</td>
<td>-0.0000 - 0.0100</td>
</tr>
<tr>
<td>EDUC</td>
<td>0.02899**</td>
<td>0.00139</td>
<td>21.91</td>
<td>0.00000</td>
<td>0.0257 - 0.0321</td>
</tr>
<tr>
<td>MARRIED</td>
<td>0.14510**</td>
<td>0.00547</td>
<td>27.06</td>
<td>0.00000</td>
<td>0.1333 - 0.1569</td>
</tr>
<tr>
<td>HSAT</td>
<td>0.07717**</td>
<td>0.00122</td>
<td>6.31</td>
<td>0.00000</td>
<td>0.0747 - 0.0796</td>
</tr>
<tr>
<td>Constant</td>
<td>0.05845</td>
<td>0.00779</td>
<td>7.70</td>
<td>0.00000</td>
<td>0.0433 - 0.0737</td>
</tr>
<tr>
<td>AGE</td>
<td>0.00395**</td>
<td>0.00416</td>
<td>0.96</td>
<td>0.33333</td>
<td>-0.0000 - 0.0100</td>
</tr>
<tr>
<td>EDUC</td>
<td>0.02899**</td>
<td>0.00139</td>
<td>21.91</td>
<td>0.00000</td>
<td>0.0257 - 0.0321</td>
</tr>
<tr>
<td>MARRIED</td>
<td>0.14510**</td>
<td>0.00547</td>
<td>27.06</td>
<td>0.00000</td>
<td>0.1333 - 0.1569</td>
</tr>
<tr>
<td>HSAT</td>
<td>0.07717**</td>
<td>0.00122</td>
<td>6.31</td>
<td>0.00000</td>
<td>0.0747 - 0.0796</td>
</tr>
<tr>
<td>Constant</td>
<td>0.05845</td>
<td>0.00779</td>
<td>7.70</td>
<td>0.00000</td>
<td>0.0433 - 0.0737</td>
</tr>
<tr>
<td>AGE</td>
<td>0.00395**</td>
<td>0.00416</td>
<td>0.96</td>
<td>0.33333</td>
<td>-0.0000 - 0.0100</td>
</tr>
<tr>
<td>EDUC</td>
<td>0.02899**</td>
<td>0.00139</td>
<td>21.91</td>
<td>0.00000</td>
<td>0.0257 - 0.0321</td>
</tr>
<tr>
<td>MARRIED</td>
<td>0.14510**</td>
<td>0.00547</td>
<td>27.06</td>
<td>0.00000</td>
<td>0.1333 - 0.1569</td>
</tr>
<tr>
<td>HSAT</td>
<td>0.07717**</td>
<td>0.00122</td>
<td>6.31</td>
<td>0.00000</td>
<td>0.0747 - 0.0796</td>
</tr>
</tbody>
</table>

**Note:** *** ** * ** Significant at 1%, 5%, 10% level.
Quantile Linear Regression: SAS - Application 3

```sas
proc quantreg data=ab;
model y1 = xm SMB HML /quantile=0.25 0.5 0.75
run;
```

The QUANTREG Procedure
Quantile and Objective Function

<table>
<thead>
<tr>
<th>Quantile</th>
<th>Parameter</th>
<th>DF</th>
<th>Estimate</th>
<th>95% Confidence Limits</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.25</td>
<td>Intercept</td>
<td>1</td>
<td>-1.6310</td>
<td>-1.7793 -1.5164</td>
</tr>
<tr>
<td></td>
<td>xm</td>
<td>1</td>
<td>0.9855</td>
<td>0.9477 1.0069</td>
</tr>
<tr>
<td></td>
<td>SMB</td>
<td>1</td>
<td>1.2018</td>
<td>1.1505 1.3219</td>
</tr>
<tr>
<td></td>
<td>HML</td>
<td>1</td>
<td>0.5071</td>
<td>0.4250 0.5615</td>
</tr>
<tr>
<td>0.75</td>
<td>Intercept</td>
<td>1</td>
<td>0.9056</td>
<td>0.6957 1.1344</td>
</tr>
<tr>
<td></td>
<td>xm</td>
<td>1</td>
<td>0.9919</td>
<td>0.9626 1.0535</td>
</tr>
<tr>
<td></td>
<td>SMB</td>
<td>1</td>
<td>1.4267</td>
<td>1.3340 1.5025</td>
</tr>
<tr>
<td></td>
<td>HML</td>
<td>1</td>
<td>0.5435</td>
<td>0.4593 0.6213</td>
</tr>
</tbody>
</table>

Heteroscedasticity

- Model: \( y_i = x_i'\beta + \epsilon_i \), with i.i.d. errors.
  - The quantiles are a vertical shift of one another.

- Model: \( y_i = x_i'\beta + \sigma(x_i) \epsilon_i \), errors are now heteroscedastic.
  - The quantiles now exhibit a location shift as well as a scale shift.

- Khmaladze-Koenker Test Statistic
Quantile Regression: Bibliography


Quantile Regression

- S+ Programs - Lib.stat.cmu.edu/s
- [www.econ.uiuc.edu/~roger](http://www.econ.uiuc.edu/~roger)
- [http://Lib.stat.cmu.edu/R/CRAN](http://Lib.stat.cmu.edu/R/CRAN)
- TSP
- Limdep