Chapter 11: Robust & Quantile regression

Timothy Hanson

Department of Statistics, University of South Carolina

Stat 704: Data Analysis I
11.3: Robust regression

- Leverages $h_{ii}$ and deleted residuals $t_i$ useful for finding outlying $x_i$ and $Y_i$ (w.r.t. the model) cases.
- Cook’s $D_i$ and DFFIT$_i$ indicate which cases are highly influencing the fit of the model, i.e. the OLS $b$.

What to do with influential and/or outlying cases? Are they transcription errors or somehow unrepresentative of the target population?

- Outliers are often interesting in their own right and can help guide the building of a better model.
- Robust regression dampens the effect of outlying cases on estimation to provide a better fit to the majority of cases.
- Useful in situations when there’s no time for “influence diagnostics” or a more careful analysis.
Robust regression is effective when the error distribution is not normal, but heavy-tailed.

M-estimation is a general class of estimation methods. Choose $\beta$ to minimize

$$Q(\beta) = \sum_{i=1}^{n} \rho(Y_i - x_i'\beta),$$

where $\rho(\cdot)$ is some function.

- $\rho(u) = u^2$ gives OLS $b$.
- $\rho(u) = |u|$ gives $L_1$ regression, or least absolute residual (LAR) regression.

Huber’s method – described next – builds a $\rho(\cdot)$ that is a compromise between OLS and LAR. It looks like $u^2$ for $u$ close to zero and $|u|$ further away.
Iteratively Reweighted Least Squares

Outlying values of $r_i^j = y_i - \mathbf{x}_i' \mathbf{b}^j$ are (iteratively) given less weight in the estimation process.

0 (Starting $\mathbf{b}^0$) OLS: $w_i^0 = 1/e_i^2$ from $\mathbf{e} = \mathbf{Y} - \mathbf{X}\mathbf{b}$. Set $j = 1$.

1 (WLS for $\mathbf{b}^j$ using $\mathbf{W}^{j-1}$): $\mathbf{b}^j = (\mathbf{X}'\mathbf{W}^{j-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}^{j-1}\mathbf{Y}$.

2 $\hat{\sigma}^j = \text{median}\{|Y_i - \mathbf{x}_i' \mathbf{b}^j|/\Phi^{-1}(0.75) : i = 1, \ldots, n\}$.

3 $w_i^j = W\left(\frac{|Y_i - \mathbf{x}_i' \mathbf{b}^j|}{\hat{\sigma}^j}\right)$, where

$$W(u) = \begin{cases} 
1 & \text{if } |u| \leq 1.345 \\
\frac{1.345}{|u|} & \text{if } |u| > 1
\end{cases}.$$

Set $j$ to $j + 1$.

1 Repeat steps 1 through 3 until $\hat{\sigma}^j$ and $\mathbf{b}^j$ stabilize.

There are other weight functions (SAS default is bisquare, pp. 439-440) and other methods for updating $\hat{\sigma}^j$ – see book and SAS documentation if interested.
11.3 Influential cases rem. measure: Robust regression
Quantile regression

SAS code: M-estimation w/ Huber weight

```sas
data prof;
  input state$ mathprof parents homelib reading tvwatch absences;
datalines;
  Alabama       252   75  78  34  18  18
  Arizona       259   75  73  41  12  26
  ...et cetera...
  Wisconsin     274   81  86  38  8   21
  Wyoming       272   85  86  43  7   23
;
proc robustreg data=prof method=m (wf=huber);
  model mathprof=parents homelib reading tvwatch absences; run;
proc robustreg data=prof method=m (wf=huber);
  model mathprof=parents homelib reading tvwatch absences;
  test absences reading; run;
proc robustreg data=prof method=m (wf=huber);
  model mathprof=parents homelib tvwatch;
  id state; output out=out p=p sr=sr; run;
goptions reset=all; symbol1 v=dot pointlabel=("#state");
proc gplot data=out; plot sr*(p parents homelib tvwatch); run;
```

### The ROBUSTREG Procedure

#### Parameter Estimates

<table>
<thead>
<tr>
<th>Parameter</th>
<th>DF</th>
<th>Estimate</th>
<th>Standard Error</th>
<th>95% Confidence Limits</th>
<th>Chi-Square</th>
<th>Pr &gt; ChiSq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1</td>
<td>178.1630</td>
<td>28.2897</td>
<td>122.7162 - 233.6097</td>
<td>39.66</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>parents</td>
<td>1</td>
<td>0.3574</td>
<td>0.2007</td>
<td>-0.0360 - 0.7507</td>
<td>3.17</td>
<td>0.0750</td>
</tr>
<tr>
<td>homelib</td>
<td>1</td>
<td>0.7702</td>
<td>0.1403</td>
<td>0.4952 - 1.0452</td>
<td>30.14</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>reading</td>
<td>1</td>
<td>0.1496</td>
<td>0.1403</td>
<td>0.0162 - 0.7085</td>
<td>36.01</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>tvwatch</td>
<td>1</td>
<td>-0.8218</td>
<td>0.2752</td>
<td>-1.3612 - 0.2824</td>
<td>8.92</td>
<td>0.0028</td>
</tr>
<tr>
<td>absences</td>
<td>1</td>
<td>-0.0121</td>
<td>0.2058</td>
<td>-0.4155 - 0.3912</td>
<td>0.00</td>
<td>0.9530</td>
</tr>
</tbody>
</table>

#### Robust Linear Tests

<table>
<thead>
<tr>
<th>Test</th>
<th>Chi-Square</th>
<th>Pr &gt; ChiSq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rho</td>
<td>0.25</td>
<td>0.8805</td>
</tr>
<tr>
<td>Rn2</td>
<td>2</td>
<td>0.51</td>
</tr>
</tbody>
</table>

#### Parameter Estimates

<table>
<thead>
<tr>
<th>Parameter</th>
<th>DF</th>
<th>Estimate</th>
<th>Standard Error</th>
<th>95% Confidence Limits</th>
<th>Chi-Square</th>
<th>Pr &gt; ChiSq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1</td>
<td>185.8796</td>
<td>21.4793</td>
<td>143.7809 - 227.9784</td>
<td>74.89</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>parents</td>
<td>1</td>
<td>0.3623</td>
<td>0.1766</td>
<td>0.0162 - 0.7085</td>
<td>4.21</td>
<td>0.0402</td>
</tr>
<tr>
<td>homelib</td>
<td>1</td>
<td>0.7543</td>
<td>0.1257</td>
<td>0.5080 - 1.0007</td>
<td>36.01</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>tvwatch</td>
<td>1</td>
<td>-0.9369</td>
<td>0.2184</td>
<td>-1.3650 - 0.5089</td>
<td>18.40</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Scale</td>
<td>1</td>
<td>4.2666</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
11.3 Influential cases rem. measure: Robust regression
Quantile regression
11.3 Influential cases rem. measure: Robust regression

Quantile regression
11.3 Influential cases rem. measure: Robust regression
Quantile regression
11.3 Influential cases rem. measure: Robust regression
Quantile regression
Quantile regression

- Another option is general *quantile regression*.

- $L_1$ regression (p. 438), described above, is *median regression* and can be carried out in SAS PROC QUANTREG. Other quantiles can similarly be regressed upon.

- The QUANTREG procedure uses robust multivariate location and scale estimates for leverage point detection. So you can do similar analyses as in PROC REG, but robustly.

- QUANTREG solves minimization problem using simplex algorithm, details in documentation.
Median, or $L_1$ regression minimizes

$$Q_{0.5}(\beta) = \min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} |Y_i - x'_i\beta|.$$  

Let $0 < \tau < 1$ be a probability. Koenker and Bassett (1978) define $b_{\tau}$, the $\tau$th quantile regression effects vector, to minimize

$$Q_{\tau}(\beta) = \min_{\beta \in \mathbb{R}^p} \left[ \tau \sum_{i:y_i \geq x'_i\beta} |Y_i - x'_i\beta| + (1 - \tau) \sum_{i:y_i < x'_i\beta} |Y_i - x'_i\beta| \right].$$

Let $Y_h$ be a draw from the population with accompanying $x_h$. The $b_{\tau}$ that minimizes $Q_{\tau}(\beta)$ satisfies

$$P(Y_h \leq x'_h b_{\tau}) \approx \tau.$$ 

Note $q_{\tau}(x) = x'b_{\tau}$. How are the elements of $b_{\tau}$ interpreted?
11.3 Influential cases rem. measure: Robust regression

Quantile regression

SAS code: Lung pressure data

data pressure;
   input spap empty eject gas @@;
datalines;
   49.0 45.0 36.0 45.0 55.0 30.0 28.0 40.0 85.0 11.0 16.0 42.0
   32.0 30.0 46.0 40.0 26.0 39.0 76.0 43.0 28.0 42.0 78.0 27.0
   95.0 17.0 24.0 36.0 26.0 63.0 80.0 42.0 74.0 25.0 12.0 52.0
   37.0 32.0 27.0 35.0 31.0 37.0 37.0 55.0 49.0 29.0 34.0 47.0
   38.0 26.0 32.0 28.0 41.0 38.0 45.0 30.0 12.0 38.0 99.0 26.0
   44.0 25.0 38.0 47.0 29.0 27.0 51.0 44.0 40.0 37.0 32.0 54.0
   31.0 34.0 40.0 36.0;

data pressure2; set pressure; eject2=eject**2;

proc quantreg data=pressure2;
   model spap=eject / quantile=0.25;
   output out=o1 p=p1;
proc quantreg data=pressure2;
   model spap=eject / quantile=0.50;
   output out=o2 p=p2;
proc quantreg data=pressure2;
   model spap=eject / quantile=0.75;
   output out=o3 p=p3;
data o4; set o1 o2 o3; proc sort data=o4; by quantile eject;
goptions reset=all;
symbol1 color=black value=dot interpol=none;
symbol2 color=black value=none l=3 interpol=join;
symbol3 color=black value=none l=1 interpol=join;
symbol4 color=black value=none l=3 interpol=join;
legend1 value=("data" "25%" "50%" "75%");
proc gplot data=o4; plot spap*eject p1*eject p2*eject p3*eject / overlay legend=legend1;
### Quantile regression

The QUANTREG Procedure

#### Quantile 0.25

- **Intercept**: 1, **DF**: 1, **Estimate**: 49.3585, **Limits**: 35.9819 - 88.2810
- **eject**: 1, **Estimate**: -0.3774, **Limits**: -2.4717 - 0.1364

#### Quantile 0.5

- **Intercept**: 1, **DF**: 1, **Estimate**: 65.1667, **Limits**: 45.3958 - 83.0584
- **eject**: 1, **Estimate**: -0.5370, **Limits**: -1.1718 - 0.2699

#### Quantile 0.75

- **Intercept**: 1, **DF**: 1, **Estimate**: 82.5517, **Limits**: 65.5742 - 110.0804
- **eject**: 1, **Estimate**: -0.7126, **Limits**: -0.9293 - 0.5127
11.3 Influential cases rem. measure: Robust regression
Quantile regression
Normal-errors regression is also quantile regression

For our standard model:

\[ Y = \mathbf{x}'\beta + \epsilon, \quad \epsilon \sim N(0, \sigma^2), \]

note that

\[
P(Y \leq q_\tau) = \tau \iff P(\mathbf{x}'\beta + \epsilon \leq q_\tau) = \tau \iff \\
P\left( \frac{\epsilon}{\sigma} \leq \frac{q_\tau - \mathbf{x}'\beta}{\sigma} \right) = \tau \iff \\
\Phi\left( \frac{q_\tau - \mathbf{x}'\beta}{\sigma} \right) = \tau \iff \\
q_\tau - \mathbf{x}'\beta = \sigma \Phi^{-1}(\tau) \iff \\
q_\tau(\mathbf{x}) = \Phi^{-1}(\tau)\sigma + \mathbf{x}'\beta.
\]
• When $x_j$ increases by one, every quantile increases by $\beta_j$.
• What does this imply about the quantile functions?
• Could we see a plot like we saw for the lung pressure data a few slides ago?
• Final comment: $L_1$ regression is obtained as the MLE from a standard regression model assuming the errors are distributed double-exponential (Laplace).