Pattern in the residuals

Consider a data set on depression...

Dependent variable:
Beck’s Depression Inventory (BDI) scale 0-63

Predictors:
income (continuous)
etnicity (factor, 6 levels)
marital (factor, 7 levels)
abortion (yes or no)
And the following classical linear model...

```r
> lm.simple=lm(bdi.score ~ income + ethnic
                   marital + abort)
```

with one continuous covariate, two categorical variables, and one factor of interest.
Residual plots:

Residual plots after log$_e$ transformation:

Normality is much better, but what’s going on with the constant variance plot?
Look at the original plot:

![Plot](image)

BDI is discrete and bounded by zero on the left.

Suppose an observation has a predicted value of 7. What is the smallest (i.e. most negative) possible residual this observation could have? \( Y - \hat{Y} = 0 - 7 = -7. \)

Similarly, the smallest possible residual for a predicted value of 10 is \( 0 - 10 = -10. \)

The residuals can not go below the line \( Y = -x. \)
If this bounding seems to suggest the conditional distributions are not close to normal (i.e. it looks VERY chopped off), then perhaps a different modeling would be a better choice.

Just knowing that the BDI was bounded between 0 and 63 does not tell us this residual pattern will be strong enough to be a problem. It depends on the actual fitted values and the actual observed values.

The problem tends to be most extreme when a large amount of the data is predicted by the linear model to be near the bound, such as $\hat{Y} = 0$.

The other question that seems relevant to the residual plot is the sparse number of predictions with $\hat{Y} > 13$. 
> table(ethnic)

<table>
<thead>
<tr>
<th>ethnic</th>
<th>5</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>natAlaskan/AmerInd</td>
<td>2367</td>
<td>39</td>
<td>106</td>
<td>155</td>
<td>134</td>
<td>40</td>
</tr>
</tbody>
</table>

> table(marital)

<table>
<thead>
<tr>
<th>marital</th>
<th>1</th>
<th>2</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>366</td>
<td>1947</td>
<td>77</td>
<td>125</td>
<td>316</td>
<td>10</td>
<td></td>
</tr>
</tbody>
</table>
Similar situation from an earlier semester

A social work client was predicting ‘Anger/Aggression’ on a discrete scale from 0 to 27:

![Histogram of AngerAggression](image)

> `lm.out=lm(AngerAggression ~ Gender + SES + Age + ... + Conflict)`
Residual plot from linear model:

In the end, a Poisson regression was fitted to the data (their suggestion). The assumptions of that model are more difficult to check than a classical linear model, but perhaps it provided a better fit to the conditional distributions.
Past client’s linguistics model

The response is the proportion correct out of 12 questions and is bounded between 0 and 1.

```
proc mixed data=combinedMod;
class Group Condition ID;
model PropCorrect = Group Condition Group*Condition/ddfm=satterth;
random ID(Group);
run;
```
Checking constant variance assumption:

Checking normality assumption:
Note only is the response discrete, but it is bounded between 0 and 1.

The residuals on the red line had observations = 1.
The residuals on the green line had observations = \(\frac{11}{12}\).
The residuals on the blue line had observations = \(\frac{10}{12}\).
Past client: Yupeng Kou

Conditional Residuals for taskscore100

Residual Statistics
- Observations: 168
- Minimum: -36.38
- Mean: 15E-15
- Maximum: 41.7
- Std Dev: 13.636

Fit Statistics
- Objective: 1259.7
- AIC: 1263.7
- AICC: 1263.8
- BIC: 1265
Ben Miller for Oucher:

```
jitter(lm.out$fitted, factor = 0.2)
lm.out$residuals
```
This is another linear model residual plot when the response was bounded between 0 to 1.

This one had a fair number of observations near both ends of the bounded range.
Modeling option: Beta regression

If your response is continuous but bounded between 0 and 1, *beta regression* is an option.

\[ Y \mid x \sim \text{beta}(\alpha(x), \beta(x)) \]
You can perform hypothesis tests on the $\alpha$’s and $\beta$’s parameters to test if groups are statistically significantly different from each other.

The observed data may fit this type of model better, but you may have difficulty convincing a client to use this type of a model (for publication submission, for instance).