Table 1: The measured gain for the middle 10 runs of 30 of the snow gauge, for each of 9 densities in grams per cubic centimeter of polyethylene blocks (USDA Forest Service).

<table>
<thead>
<tr>
<th>Density</th>
<th>Gain</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.686</td>
<td>17.6 17.3 16.9 16.2 17.1 18.5 18.7 17.4 18.6 16.8</td>
</tr>
<tr>
<td>0.604</td>
<td>24.8 25.9 26.3 24.8 24.8 27.6 28.5 30.5 28.4 27.7</td>
</tr>
<tr>
<td>0.508</td>
<td>39.4 37.6 38.1 37.7 36.3 38.7 39.4 38.8 39.2 40.3</td>
</tr>
<tr>
<td>0.412</td>
<td>60.0 58.3 59.6 59.1 56.3 55.0 52.9 54.1 56.9 56.0</td>
</tr>
<tr>
<td>0.412</td>
<td>39.4 37.6 38.1 37.7 36.3 38.7 39.4 38.8 39.2 40.3</td>
</tr>
<tr>
<td>0.412</td>
<td>60.0 58.3 59.6 59.1 56.3 55.0 52.9 54.1 56.9 56.0</td>
</tr>
<tr>
<td>0.318</td>
<td>87.0 92.7 90.5 85.8 87.5 88.3 91.6 88.2 88.6 84.7</td>
</tr>
<tr>
<td>0.223</td>
<td>128 130 131 129 127 129 132 133 134 133</td>
</tr>
<tr>
<td>0.148</td>
<td>199 204 199 207 200 200 205 202 199 199</td>
</tr>
<tr>
<td>0.080</td>
<td>298 298 297 288 296 293 301 299 298 293</td>
</tr>
<tr>
<td>0.001</td>
<td>423 421 422 428 436 427 426 428 427 429</td>
</tr>
</tbody>
</table>

Replicate Measurements

Notation

• $x$ represents:

• $y$ represents:

• $n =$

• Notation is simpler if we put double subscripts on $y$
  
  – Let $i$ represent the level of density. So $i = 1, \ldots$
  
  – Let $j$ represent the replicate measurement of gain. So $j = 1, \ldots$

• The $y_{1,1} = \ldots$, $y_{3,1} = \ldots$ and $y_{2,9} = \ldots$

• Also, let $\bar{y}_i$ represent the average of the replicates for group (density) $i$.

\[
\bar{y}_i = \frac{1}{?} \sum_{j=1}^{?} y,
\]
The Model

Note that here we are preselecting the density values. They are design points for our experiment. This means that the $x$s are nonrandom.

We don’t need a double subscript for the $xx$ – why?

$x_1 = _____$, $x_3 = _____$, and $x_7 = _____$. In general, $x_i$ represents density for $i = 1, \ldots$

The model for gain is then:

$$Y_{i,j} = a + bx_i + E_{i,j}$$

where the errors are mean 0 and constant variance for all $i, j$.

Fitting by least squares

The least squares fit of $a$ and $b$ is found by minimizing the following sums of squares:

$$\sum \sum (y_{i,j} - (a + bx_i))^2$$

The minimizing $a$ and $b$ are the same if we minimize

$$\sum \sum (\bar{y}_i - (a + bx_i))^2$$

PROVE THIS
Errors

So where does the difference between fitting these two models enter the picture?
Are the predicted values the same or different? Explain

Are the residuals the same or different? Explain

The residuals from the first approach can be decomposed into two parts – one which can be used to assess model misfit and the other can be used to estimate the variance of $E_{i,j}$.

$$\sum_i \sum_j (y_{i,j} - \hat{y}_{i,j})^2 = \sum_i \sum_j (y_{i,j} - \cdots - \hat{y}_{i,j})^2$$

$$= \sum_i \sum_j (y_{i,j} - \cdots)^2 + (\cdots - \hat{y}_{i,j})^2$$

Establish that the cross product term is 0.

Which of the two terms is measures model misfit and which measures the variance of $E_{i,j}$? Explain