Section 4.1: Fitting a Line by Least Squares

Often we want to fit a straight line to data.

For example from an experiment we might have the following data showing the relationship of density of specimens made from a ceramic compound at different pressures.

By fitting a line to the data we can predict what the average density would be for specimens made at any given temperature, even pressures we did not investigate experimentally.

For a straight line we assume a model which says that on average in the whole population of possible specimens the average density, $y$, value is related to pressure, $x$, by the equation

$$y \approx \beta_0 + \beta_1 x$$

The population (true) intercept and slope are represented by Greek symbols just like $\mu$ and $\sigma$. 
For the measured data we fit a straight line

\[ \hat{y} = b_0 + b_1 x \]

For the \( i \)th point, the fitted line or predicted value is

\[ \hat{y}_i = b_0 + b_1 x_i \]

The fitted line is most often determined by the method of “least squares”.

This is the optimal method to use for fitting the line if

- The relationship is in fact linear.
- For a fixed value of \( x \) the resulting values of \( y \) are
  - normally distributed with
  - the same constant variance at all \( x \) values.

If these assumptions are not met, then we are not using the best tool for the job.

For any statistical tool, know when that tool is the right one to use.
A least squares fit minimizes the sum of squared deviations from the fitted line \( \hat{y} \)

\[
\text{minimize } \sum (y_i - \hat{y}_i)^2
\]

Deviations from the fitted line are called “residuals”
- We are minimizing the sum of squared residuals,
- called the “residual sum of squares.”

We need to
- minimize \( \sum (y_i - (b_0 + b_1 x_i))^2 \)
- over all possible values of \( b_0 \) and \( b_1 \)
- a calculus problem.

The resulting formulas for the least squares estimates of the intercept and slope are

\[
b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}
\]

\[
b_0 = \bar{y} - b_1 \bar{x}
\]

\[
\hat{y} = \bar{y} - b_1 \bar{x} + b_1 x
\]

\[
\hat{y} - \bar{y} = b_1 (x - \bar{x})
\]
\hat{y} - \bar{y} = b_i(x - \bar{x}). \text{ When } x = \bar{x}, \text{ then } y = \bar{y}.

If we have average pressure, \(x\), then we expect to get about average density, \(y\).

The sample (linear) correlation coefficient, \(r\), is a measure of how “correlated” the \(x\) and \(y\) variable are.

The correlation coefficient is between -1 and 1

+1 means perfectly positively correlated
0 means no correlation
-1 means perfectly negatively correlated

The correlation coefficient is computed by

\[
r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}
\]
The slope, \( b_1 \), and the correlation coefficient are related by

\[
b_1 = \frac{r \cdot SD_y}{SD_x} = \frac{\text{rise}}{\text{run}}
\]

- \( SD_y \) is the standard deviation of y values and
- \( SD_x \) is the standard deviation of x values.

For every \( SD_x \) run on the x axis, the fitted line rises \( r \cdot SD_y \) units on the y axis.

So if x is a certain number of standard deviations above average, \( \bar{x} \),

- then y is, on average, the fraction r times that many standard deviations above average, \( \bar{y} \).

For example if

- the correlation is \( r = 0.9 \)
- the pressure, x, is 2 standard deviations above average \( \bar{x} \)
- Then we expect the density, y, will be about
  
  \[ 0.9(2) = 1.8 \text{ standard deviations above average } \bar{y}. \]
Interpretation of $r^2$ or $R^2$

$R^2 =$ fraction of variation accounted for (explained by) the fitted line.

Ceramic Items Page 124

<table>
<thead>
<tr>
<th>Pressure</th>
<th>y = Density</th>
<th>y - mean</th>
<th>(y-mean)^2</th>
</tr>
</thead>
<tbody>
<tr>
<td>2000</td>
<td>2486</td>
<td>-181</td>
<td>32761</td>
</tr>
<tr>
<td>2000</td>
<td>2479</td>
<td>-188</td>
<td>35344</td>
</tr>
<tr>
<td>2000</td>
<td>2472</td>
<td>-195</td>
<td>38025</td>
</tr>
<tr>
<td>4000</td>
<td>2558</td>
<td>-109</td>
<td>11881</td>
</tr>
<tr>
<td>4000</td>
<td>2570</td>
<td>-97</td>
<td>9409</td>
</tr>
<tr>
<td>4000</td>
<td>2580</td>
<td>-87</td>
<td>7569</td>
</tr>
<tr>
<td>6000</td>
<td>2646</td>
<td>-21</td>
<td>441</td>
</tr>
<tr>
<td>6000</td>
<td>2657</td>
<td>-10</td>
<td>100</td>
</tr>
<tr>
<td>6000</td>
<td>2653</td>
<td>-14</td>
<td>196</td>
</tr>
<tr>
<td>8000</td>
<td>2724</td>
<td>57</td>
<td>3249</td>
</tr>
<tr>
<td>8000</td>
<td>2774</td>
<td>107</td>
<td>11449</td>
</tr>
<tr>
<td>8000</td>
<td>2808</td>
<td>141</td>
<td>19881</td>
</tr>
<tr>
<td>10000</td>
<td>2861</td>
<td>194</td>
<td>37636</td>
</tr>
<tr>
<td>10000</td>
<td>2879</td>
<td>212</td>
<td>44944</td>
</tr>
<tr>
<td>10000</td>
<td>2858</td>
<td>191</td>
<td>36481</td>
</tr>
</tbody>
</table>

mean 6000 2667 sum 0 289366
st dev 2927.7 143.767
correlation 0.991
correl^2 0.982
- If we don't use the pressures to predict density
  - We use $\bar{y}$ to predict every $y_i$
  - Our sum of squared errors is $\sum (y_i - \bar{y})^2 = 289,366 =$ SS Total in Excel
- If we do use the pressures to predict density
  - We use $\hat{y}_i = b_0 + b_1 x_i$ to predict $y_i$
  - $\sum (y_i - \hat{y}_i)^2 = 5152.67 =$ SS Residual in Excel

<table>
<thead>
<tr>
<th>Observation</th>
<th>Predicted Density</th>
<th>Residuals</th>
<th>Residual^2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2472.333</td>
<td>13.667</td>
<td>186.778</td>
</tr>
<tr>
<td>2</td>
<td>2472.333</td>
<td>6.667</td>
<td>44.444</td>
</tr>
<tr>
<td>3</td>
<td>2472.333</td>
<td>-0.333</td>
<td>0.111</td>
</tr>
<tr>
<td>4</td>
<td>2569.667</td>
<td>-11.667</td>
<td>136.111</td>
</tr>
<tr>
<td>5</td>
<td>2569.667</td>
<td>0.333</td>
<td>0.111</td>
</tr>
<tr>
<td>6</td>
<td>2569.667</td>
<td>10.333</td>
<td>106.778</td>
</tr>
<tr>
<td>7</td>
<td>2667.000</td>
<td>-21.000</td>
<td>441.000</td>
</tr>
<tr>
<td>8</td>
<td>2667.000</td>
<td>-10.000</td>
<td>100.000</td>
</tr>
<tr>
<td>9</td>
<td>2667.000</td>
<td>-14.000</td>
<td>196.000</td>
</tr>
<tr>
<td>10</td>
<td>2764.333</td>
<td>-40.333</td>
<td>1626.778</td>
</tr>
<tr>
<td>11</td>
<td>2764.333</td>
<td>9.667</td>
<td>93.444</td>
</tr>
<tr>
<td>12</td>
<td>2764.333</td>
<td>43.667</td>
<td>1906.778</td>
</tr>
<tr>
<td>13</td>
<td>2861.667</td>
<td>-0.667</td>
<td>0.444</td>
</tr>
<tr>
<td>14</td>
<td>2861.667</td>
<td>17.333</td>
<td>300.444</td>
</tr>
<tr>
<td>15</td>
<td>2861.667</td>
<td>-3.667</td>
<td>13.444</td>
</tr>
</tbody>
</table>

5152.666667 sum
The percent reduction is our error sum of squares is

\[ R^2 = \frac{\sum (y_i - \bar{y})^2 - \sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2} \times 100 = \frac{289,366 - 512.67}{289,366} \times 100 = \frac{284,213.33}{289,366} \times 100 \]

\[ R^2 = 98.2\% \]

Using x to predict y decreases the error sum of squares by 98.2%.

The reduction in error sum of squares from using x to predict y is

- Sum of squares explained by the regression equation
- 284,213.33 = SS Regression in Excel

This is also the correlation squared.

\[ r^2 = 0.991^2 = 0.982 \]

For a perfectly straight line

- All residuals are zero.
  - The line fits the points exactly.
- SS Residual = 0
- SS Regression = SS Total
  - The regression equation explains all variation
- \( R^2 = 100\% \)
- \( r = \pm 1 \)
  - \( r^2 = 1 \)

If \( r=0 \), then there is no linear relationship between x and y

- \( R^2 = 0\% \)
- Using x to predict y does not help at all.
Checking Model Adequacy

With only single x variable, we can tell most of what we need from a plot with the fitted line.

Plotting residuals will be most crucial in section 4.2 with multiple x variables
  • But residual plots are still of use here.

Plot residuals versus
  • Predicted values \( \hat{y} \)
  • Versus x
  • In run order
  • Versus other potentially influential variables, e.g. technician
  • Normal Plot of residuals
A residual plot gives us a magnified view of the increasing variance and curvature.

This residual plot indicates 2 problems with this linear least squares fit
- The relationship is not linear
  - Indicated by the curvature in the residual plot
- The variance is not constant
  - So the least squares method isn't the best approach even if we handle the nonlinearity.

Don't fit an exponential function to these data directly with least squares.
- With increasing variability, not all squared errors should count equally.
  - An error of 5 for x=1
  - Is a bigger problem than an error of 5 for x=16.
Some Study Questions

What does it mean to say that a line fit to data is the "least squares" line? Where do the terms least and squares come from?

We are fitting data with a straight line. What 3 assumptions (conditions) need to be true for a linear least squares fit to be the optimal way of fitting a line to the data?

What does it mean if the correlation between x and y is -1? What is the residual sum of squares in this situation?

If the correlation between x and y is 0, what is the regression sum of squares, SS Regression, in this situation?

If x is 2 standard deviations above the average x value and the correlation between x and y is -0.6, the expected corresponding y values is how many standard deviations above or below the average y value?
Consider the following data.

ANOVA

<table>
<thead>
<tr>
<th></th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>Significance F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>57.6</td>
<td>57.6</td>
<td>12</td>
<td>0.041</td>
</tr>
<tr>
<td>Residual</td>
<td>3</td>
<td>14.4</td>
<td>4.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>4</td>
<td>72</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Coefficients

<table>
<thead>
<tr>
<th></th>
<th>Standard Error</th>
<th>t Stat</th>
<th>P-value</th>
<th>Lower 95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>0.8</td>
<td>0.348</td>
<td>0.751</td>
<td>-6.513</td>
</tr>
<tr>
<td>X Variable 1</td>
<td>1.2</td>
<td>3.464</td>
<td>0.041</td>
<td>0.098</td>
</tr>
</tbody>
</table>

What is the value of $R^2$?
What is the least squares regression equation?
How much does $y$ increase on average if $x$ is increased by 1.0?
What is the sum of squared residuals? Do not compute the residuals; find the answer is the Excel output.
What is the sum of squares of deviations of $y$ from $\bar{y}$?
By how much is the sum of squared errors reduced by using $x$ to predict $y$ compared to using only $\bar{y}$ to predict $y$?
What is the residual for the point with $x = 2$?