Basic idea: Minimize how far off we are when we use the line to predict \( y \) by comparing to actual \( y \).

For each individual in the data
\[
\text{Residual} = y - \hat{y} = \text{observed} \ y - \text{predicted} \ y
\]

Definition: The least squares regression line is the line that minimizes the sum of the squared residuals for all points in the dataset. The sum of squared errors = SSE is that minimum sum.

See picture on next page.

R does the work for you!

You will learn how to do this in discussion. The results look like this:

```r
lm(formula = Height ~ AvgHt, data = UCDavisM)
```

Residuals:

<table>
<thead>
<tr>
<th>Min</th>
<th>1Q</th>
<th>Median</th>
<th>3Q</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>-5.4768</td>
<td>-1.3305</td>
<td>-0.2858</td>
<td>1.2427</td>
<td>5.7142</td>
</tr>
</tbody>
</table>

Coefficients:

| Estimate | Std. Error | t value | Pr(>|t|) |
|----------|------------|---------|---------|
| (Intercept) | 16.3001 | 6.3188 | 2.580 | 0.0120 *
| AvgHt     | 0.8089    | 0.0954 | 8.479 | 2.16e-12 *** |

Residual standard error: 2.304 on 71 degrees of freedom

Example 1:
This picture shows the residuals for 4 of the individuals. The blue line comes closer to all of the points than any other line, where “close” is defined by
\[
\text{SSE} = \sum_{\text{all values}} \text{residual}^2
\]

\( \text{SSE} = 376.9 \) (average of about 5.16 per person, or about 2.25 inches when take square root)

Example 2: A NEGATIVE ASSOCIATION

- A study was done to see if the distance at which drivers could read a highway sign at night changes with age.
- Data consist of \( n = 30 \) \((x, y)\) pairs where \( x = \text{Age} \) and \( y = \text{Distance at which the sign could first be read (in feet)} \).

The regression equation is
\[
\hat{y} = 577 - 3x
\]

Notice negative slope
Ex: \( 577 - 3(20) = 577 - 60 = 517 \)

<table>
<thead>
<tr>
<th>Age</th>
<th>Pred. distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>20 years</td>
<td>517 feet</td>
</tr>
<tr>
<td>50 years</td>
<td>427 feet</td>
</tr>
<tr>
<td>80 years</td>
<td>337 feet</td>
</tr>
</tbody>
</table>

Interpretation of slope and intercept?
Not easy to find the best line by eye!

Applets:

http://www.rossmanchance.com/applets/RegShuffle.htm

(Try copying and pasting data from other examples.)

http://illuminations.nctm.org/Activity.aspx?id=4187

MORE ABOUT THE MODEL: CONDITIONS and ASSUMPTIONS
(Next time we will learn how to check and correct these, in the “ASSESS” step)

1. Linearity: The linear model says is that a straight line is appropriate.
2. The variance (standard deviation) of the Y-values is constant for all values of X in the range of the data.
3. Independence: The errors are independent of each other, so knowing the value of one doesn’t help with the others.

Sometimes, also require:
4. Normality assumption: The errors are normally distributed
5. Random or representative sample, if we want to extend the results to the population.

Example 3: Predicting final average from midterm
- Relationship is linear, positive association
- Regression equation: \( \hat{y} = 46.45 + 0.4744x \) (Interpretation?)
- For instance, here are predictions for \( x = 80, 50, 100 \)
  Midterm = 80, predicted avg = 46.45 + 0.4744(80) = 84.4
  \( x = 50 \), \( \hat{y} = 70.17 \)
  \( x = 100 \), \( \hat{y} = 93.9 \)

Putting this all together, the Simple Linear Regression Model (for the Population) is:

\[ Y = \beta_0 + \beta_1X + \varepsilon \]

where \( \varepsilon \sim N(0, \sigma) \) and all independent and \( \sigma = \) standard deviation errors = standard deviation of all Y values at each X value

Picture of this model:
Another part of the FIT: Estimating $\sigma$

- Use the residuals to estimate $\sigma$.
- Call the estimate the regression standard error

$$s = \hat{\sigma}_c = \sqrt{\frac{\text{Sum of Squared Residuals}}{n-2}} = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n-2}} = \sqrt{\text{MSE}}$$

NOTE: Degrees of freedom = $n - 2$

Example: Highway Sign Distance

$$s = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{69334}{28}} = 49.76 \text{ feet}$$

Interpretation:
At each age, $X$, there is a distribution of possible distances ($Y$) at which sign can be read. The mean is estimated to be

$$\hat{y} = 577 - 3x$$

The standard deviation is estimated to be about 50 feet.

For instance, for everyone who is 30 years old, the distribution of sign-reading distances has approximately:
Mean = $577 - 90 = 487$ feet and st. dev. = 50 feet.

See picture on white board.
For Ex 3 (grades), $s = 5$. Interpretation?