2.2 The Least-Squares Line

A ___________ line is one that describes how a dependent variable, $y$, changes as an independent variable, $x$, changes in a data set $(x_1, y_1), \cdots, (x_n, y_n)$. We use it to predict $y$ for a given $x$.

Here’s a plot of a small bivariate data set: ···

Notation includes:

- $y = \hat{\beta}_0 + \hat{\beta}_1 x$: least-squares regression line, where
  - $x$: ________________ variable
  - $y$: dependent variable
  - $\hat{\beta}_0$: $y$-intercept
  - $\hat{\beta}_1$: ________________
- $(x_i, y_i)$: $i^{th}$ data point
- $\hat{y}_i$: ________________ value of $y$ given $x = x_i$: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$
- $e_i = y_i - \hat{y}_i$: residual, the (vertical) difference between observed $y_i$ and predicted $\hat{y}_i$

We predict $y$ from $x$, so minimize vertical error in the “least squares” sense by minimizing a “sum of squared errors”

$$SSE = \sum e_i^2 = \sum (y_i - \hat{y}_i)^2 = \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

(Alas, really it should be called a “sum of squared ______________.”) Ten lines of calculus gives:
For the data set \((x_1, y_1), \ldots, (x_n, y_n)\), the least-squares line is \(y = \hat{\beta}_0 + \hat{\beta}_1 x\), where

\[
\hat{\beta}_1 = \frac{s_y}{s_x} \text{ (slope)}
\]

\[
\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \text{ (y-intercept)}
\]

e.g. Here again are data (from §2.1 notes) for 13 sparrowhawk colonies relating the % of adults in a colony that return from the previous year and the number of new adults that join the colony:

\[
x = \% \text{ Returning adults} \quad 74 \ 66 \ 81 \ 52 \ 73 \ 62 \ 52 \ 45 \ 62 \ 46 \ 60 \ 46 \ 38
\]

\[
y = \# \text{ New adults} \quad 5 \ 6 \ 8 \ 11 \ 12 \ 15 \ 16 \ 17 \ 18 \ 18 \ 19 \ 20 \ 20
\]

Use a calculator to find the least-squares line:

\[
\bar{x} =
\bar{y} =
\]

\[
s_x =
\]

\[
s_y =
\]

\[
r =
\]

\[
\Rightarrow
\hat{\beta}_1 =
\hat{\beta}_0 =
\]

So our model is \(y = \)

Or we can do it more directly. (Figure out your calculator’s labels.)

e.g. Predict the number of new adults in a colony to which 60% of last year’s adults return.

\[
\hat{y} =
\]

(Note that this is far from the data set value, \((60, \_\_\_\_\_\_)\).)
2.3 Features and Limitations of the Least-Squares Line

Properties of Least-Squares Line

- Write line in point-slope form, \( y - y_0 = m(x - x_0) \),

to see that it passes through ____________.

- The slope, \( \hat{\beta}_1 = \frac{s_y}{s_x} \), indicates that a change of ____________ in \( x \) corresponds to a change of ____________ in \( y \).

- The distinction between \( x \) and \( y \) matters because we minimized error in ____________.

- Compare variation in data to variation in modeled values and errors by considering three sums of squares:
  
<table>
<thead>
<tr>
<th>Sum of squares</th>
<th>Definition</th>
<th>Measures spread of</th>
</tr>
</thead>
<tbody>
<tr>
<td>total</td>
<td>( \sum (y_i - \bar{y})^2 )</td>
<td>data about ____________</td>
</tr>
<tr>
<td>regression (or</td>
<td>( \sum (\hat{y}_i - \bar{y})^2 )</td>
<td>predictions about ____________</td>
</tr>
<tr>
<td>error</td>
<td>( \sum (y_i - \hat{y}_i)^2 )</td>
<td>data about ____________</td>
</tr>
</tbody>
</table>

Starting from \( y_i - \bar{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y}) \) and the solutions for \( \hat{\beta}_0 \) and \( \hat{\beta}_1 \), ten lines of arithmetic gives the identity \( \sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2 \).

The coefficient of determination, \( r^2 \), measures the goodness-of-fit of the model to the data and can be understood as

\[
r^2 = \frac{\sum (y_i - \bar{y})^2 - \sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2} = \frac{\text{regression sum of squares}}{\text{total sum of squares}},
\]

\[
= \frac{\text{proportion of variance in } y \text{ explained by regression line } (r^2 \in [0, 1])}
\]
Cautions

- Don’t use least-squares line to model ________________ data.
- Don’t extrapolate (even for linear-looking data).
  
  e.g. · · ·

- Check scatterplot for ____________. Find lines with and without outlier. If they differ much, the outlier is influential \( \Rightarrow \) report ________________
  
  e.g. Adding the outlier \((60, 0)\) to sparrowhawk data changes the line from \(y = 31.93 - 0.304x\) to ________________·
  
  e.g. Adding the outlier \((0, 0)\) to sparrowhawk data changes the line from \(y = 31.93 - 0.304x\) to ________________·

- Correlation does not imply ________________

Example

e.g. (§2.2 #7) Here are data on the effect of an additive on paint drying time:

<table>
<thead>
<tr>
<th>(x) = Additive concentration (%)</th>
<th>4.0</th>
<th>4.2</th>
<th>4.4</th>
<th>4.6</th>
<th>4.8</th>
<th>5.0</th>
<th>5.2</th>
<th>5.4</th>
<th>5.6</th>
<th>5.8</th>
</tr>
</thead>
<tbody>
<tr>
<td>(y) = Drying time (hours)</td>
<td>8.7</td>
<td>8.8</td>
<td>8.3</td>
<td>8.7</td>
<td>8.1</td>
<td>8.0</td>
<td>7.7</td>
<td>7.5</td>
<td>7.2</td>
<td></td>
</tr>
<tr>
<td>Fitted, ( \hat{y} )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Residual, ( y - \hat{y} )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a. Make a scatterplot.
b. Find least-squares line.
c. Find fitted value and residual for each point.
d. If concentration is increased by .1%, by how much will drying time change?
e. Predict drying time for concentration = 4.4%.
f. For what concentration would you predict a drying time of 8.2 hours?