11 Regression

- The Correlation Coefficient
- The Least-Squares Regression Line
- Features and Limitations of the Least-Squares Line
- Inference in Regression

The Correlation Coefficient

Introduction

A *bivariate* data set consists of \( n \) \((x_1, y_1), \ldots, (x_n, y_n)\).

A *scatterplot* is a *scatterplot* of a bivariate data set.

e.g. Here are data for 13 sparrowhawk colonies relating the % of adult sparrowhawks in a colony that return from the previous year and the number of new adults that join the colony:

<table>
<thead>
<tr>
<th>%Returning adults</th>
<th>74</th>
<th>66</th>
<th>81</th>
<th>52</th>
<th>73</th>
<th>62</th>
<th>52</th>
<th>45</th>
<th>62</th>
<th>46</th>
<th>60</th>
<th>46</th>
<th>38</th>
</tr>
</thead>
<tbody>
<tr>
<td>#New adults</td>
<td>5</td>
<td>6</td>
<td>8</td>
<td>11</td>
<td>12</td>
<td>15</td>
<td>16</td>
<td>17</td>
<td>18</td>
<td>18</td>
<td>19</td>
<td>19</td>
<td>20</td>
</tr>
</tbody>
</table>

The right-hand scatterplot, below, is from these data. It shows ···
The Correlation Coefficient

The correlation coefficient, \( r \), measures the \( \quad \)and \( \quad \) of the linear relationship (if any) between \( x \) and \( y \):

\[
r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}
\]

\[
= \frac{1}{n-1} \sum_{i=1}^{n} \left( \frac{x_i - \bar{x}}{s_x} \right) \left( \frac{y_i - \bar{y}}{s_y} \right) \quad \text{(a form I prefer)}
\]

An Informal Explanation of \( r \)

- Start with a scatterplot.
- Shift reference point to \( \quad \) by subtracting \( \bar{x} \) from each \( x_i \) and \( \bar{y} \) from each \( y_i \).
- Rescale the \( x \)-axis by dividing each \( x \) coordinate by \( \quad \), and rescale the \( y \)-axis by dividing each \( y \) coordinate by \( s_y \).

Now \( x \) coordinates, \( \frac{x_i - \bar{x}}{s_x} \), have mean \( \quad \) and standard deviation \( \quad \). \( y \) coordinates, \( \frac{y_i - \bar{y}}{s_y} \), have the same mean and standard deviation.

- Analyze the sign of the \( i^{th} \) term in the last sum above, \( \left( \frac{x_i - \bar{x}}{s_x} \right) \left( \frac{y_i - \bar{y}}{s_y} \right) \), by quadrant:

\[
\begin{array}{c}
\begin{array}{c}
\text{Random } x \text{ from } [0,9), \text{ standardized}
\end{array}
\end{array}
\]

\[
\begin{array}{c}
\begin{array}{c}
\text{Random } y \text{ from } [0,9), \text{ standardized}
\end{array}
\end{array}
\]

\[
\begin{array}{c}
\begin{array}{c}
50 \text{ Random Points, standardized (} r = 0.09) \n\end{array}
\end{array}
\]

\[
\begin{array}{c}
\begin{array}{c}
13 \text{ Sparrowhawk Colonies, standardized (} r = -0.75) \n\end{array}
\end{array}
\]

e.g. For the sparrowhawk data, \( r = \quad \). For the random data, \( r = \quad \).
Properties of \( r \)

- \(-1 \leq r \leq 1\), and

\[ r = \pm 1 \implies \text{data are } \underline{\text{---------}}: \quad r \approx \pm 1 \implies \text{data are } \underline{\text{-------------}} \]

\[ r \neq 0 \implies \text{some linear relationship: } x \text{ and } y \text{ are } \textit{correlated} \]

\[ r > 0 \implies \text{slope of line is } \underline{\text{---------}} \]

\[ r < 0 \implies \text{slope of line is } \underline{\text{---------}} \]

\[ r \approx 0 \implies \text{no linear relationship: } x \text{ and } y \text{ are } \underline{\text{-------------}} \]

- \( r \) doesn’t distinguish between _____ and _____

- \( r \) doesn’t depend on _________ or ___________
Cautions

- $r$ measures strength of a linear relationship; check scatterplot to avoid using $r$ for a ___________

  e.g. The data $\{(-2, 4), (-1, 1), (0, 0), (1, 1), (2, 4)\}$ fit ___________, but $r = 0$ because the data have no ___________ relationship (draw).
  
  e.g. (from http://en.wikipedia.org/wiki/Pearson_product-moment_correlation_coefficient)

- $r$ is not resistant to the influence of ____________; don’t use it for a data set with ___________

  e.g. Adding $(0, 0)$ to the sparrowhawk data changes $r$ to ____________.

- Correlation does not imply causation:
  
  A ________________ (or lurking) variable is one ____________ under consideration that correlates with both the independent and dependent variables of interest.

  e.g.
  
  - Increasing ice cream sales are correlated with increasing ____________ rates. Does ice cream cause ____________? ______
    
    The confounding variable is ____________________________.
  
  - Sleeping with shoes on is correlated with ____________________________.
    
    Does sleeping with shoes on cause ____________? ______
    
    The confounding variable is ____________________________.

  If either the independent variable under study, or a __________ confounding variable, affects the dependent variable, then both will seem to by the (__________) criterion of correlation.

    __ cartoon
The Least-Squares Regression Line

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

The least-squares regression line is the line that the data (according to a reasonable criterion).

*Example*: Collect and plot a SRS of students’ heights \( (x) \) and weights \( (y) \).

- How many lines could we fit?
- Why aren’t data on a line?
- Estimate intercept and slope. Units?

Notation includes:

- \( y_i = \beta_0 + \beta_1 x + \epsilon_i \): an unknown true (model) regression line, where \( \beta_0 \) is the \( y \)-intercept, \( \beta_1 \) is the slope, and \( \epsilon_i \) is the \( i \)th random error

- \( y = \hat{\beta}_0 + \hat{\beta}_1 x \): estimated regression line, where
  - \( x \): ___________ variable
  - \( y \): dependent variable
  - \( \hat{\beta}_0 \): estimated \( y \)-intercept
  - \( \hat{\beta}_1 \): estimated ___________

- \( (x_i, y_i) \): \( i \)th data point

- \( \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \): ___________ value of \( y \) given \( x = x_i \):

- \( e_i = y_i - \hat{y}_i \): residual, the difference between observed \( y_i \) and predicted \( \hat{y}_i \); estimates \( \epsilon_i \)

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

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

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

e.g. Here again are data 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:

\[
\begin{align*}
    x &= \% \text{Returning adults} & 74 & 66 & 81 & 52 & 73 & 62 & 52 & 45 & 62 & 46 & 60 & 46 & 38 \\
    y &= \# \text{New adults} & 5 & 6 & 8 & 11 & 12 & 15 & 16 & 17 & 18 & 18 & 19 & 20 & 20
\end{align*}
\]

Use a calculator to find the least-squares line:

\[
\begin{align*}
    \bar{x} = \quad \bar{y} = \\
    s_x = \quad s_y = \\
    r = \\
\end{align*}
\]

\[
\Rightarrow
\begin{align*}
    \hat{\beta}_1 = \quad \hat{\beta}_0 = \\
\end{align*}
\]

So our model is \(y = \)

Or we can do it more directly. (Figure out your \underline{___________} labels.)

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

\[\hat{y} = \underline{___________}\]

(Note that this is far from the data set value, \((60, \underline{_______})\).)
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 to see that it passes through ____________.
- The slope, \( \hat{\beta}_1 = \frac{\sum xy}{\sum x^2} \), 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>( \text{SST} = \sum (y_i - \bar{y})^2 )</td>
<td>data about ________________</td>
</tr>
<tr>
<td>regression</td>
<td>( \text{SSR} = \sum (\hat{y}_i - \bar{y})^2 )</td>
<td>predictions about ________________</td>
</tr>
<tr>
<td>error</td>
<td>( \text{SSE} = \sum (y_i - \hat{y}_i)^2 )</td>
<td>data about ________________</td>
</tr>
</tbody>
</table>

  Starting from \( y_i - \bar{y} = (y_i - \bar{y}_i) + (\bar{y}_i - \bar{y}) \) and the solutions for \( \hat{\beta}_0 \) and \( \hat{\beta}_1 \), ten lines of arithmetic gives \( \text{SST} = \text{SSR} + \text{SSE} \). (Recall the __________ identity.)

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

  \[
  R^2 = \frac{\text{SST} - \text{SSE}}{\text{SST}} = \frac{\text{SSR}}{\text{SST}} = \text{proportion of variability in } y \text{ explained by regression line } (R^2 \in [0, 1])
  \]

Cautions

- Don’t use least-squares line to model ________________ data.
• To extrapolate is to make a prediction $\hat{y}$ (for $y$) from an $x$ outside the range of $x$ in the 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 $\implies$ 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 ____________________.

### Inference in Regression

#### Random Variation in the Least-Squares Estimates

Our linear model is $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, where $\beta_0$ and $\beta_1$ are unknown and $\varepsilon_i$ is the (random, unexplained) error of the $i^{th}$ measurement.

$\hat{\beta}_0$ and $\hat{\beta}_1$, as estimates of $\beta_0$ and $\beta_1$ that depend on the data, are _____________________. Similarly, $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ a random variable for each value of $x$. (Maybe we should write _______ instead of just $\hat{y}$ to emphasize that there is a different $\hat{y}$ for each _______.)

#### Simplifying Assumptions for Errors in Linear Models

Assume the linear model is correct and the errors $\varepsilon_1, \cdots, \varepsilon_n$

1. are random and independent

2. all have mean _______

3. all have the same variance _______

4. are normally distributed (so $\varepsilon_i \sim N(_______, _______)$)

Evaluate assumptions with a plot of residuals vs. fitted values and a QQ plot of residuals:
Estimate $\varepsilon_i$ by the residual $e_i = \ldots$, and estimate $\sigma^2 = \sigma^2$ from the residuals as

$$\sigma^2 \approx s^2 = \frac{1}{n-2} \sum_{i=1}^{n} (e_i - 0)^2 = \frac{1}{n-2} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \frac{\text{SSE}}{n-2}$$

(Divide by $n-2$ because ____ degrees of freedom are lost in estimating ____ and ____ from data.)

An equivalent expression (easier to find without software) is

$$s^2 = (1 - r^2) \frac{n-1}{n-2} s_y^2$$

e.g. Find the error standard deviation estimate $s$, given that we found these numbers earlier:

<table>
<thead>
<tr>
<th>$n$</th>
<th>$\bar{x}$</th>
<th>$s_x$</th>
<th>$s_y$</th>
<th>$\hat{\beta}_0$</th>
<th>$\hat{\beta}_1$</th>
<th>$r$</th>
<th>$s$</th>
</tr>
</thead>
<tbody>
<tr>
<td>13</td>
<td>58.23</td>
<td>13.03</td>
<td>5.29</td>
<td>31.93</td>
<td>-0.3040</td>
<td>-.7485</td>
<td></td>
</tr>
</tbody>
</table>

Now we consider three forms of inference:

- on the slope, $\beta_1$, especially for $H_0 : \beta_1 = \ldots$

- on the $y$-intercept, $\beta_0$

- on the mean response, $E(y) = \mu_y = \beta_0 + \beta_1 x$, for a given $x$
Inference on the slope, $\beta_1$

Experts say:

- Confidence interval for $\beta_1$: $\hat{\beta}_1 \pm t_{n-2,\alpha/2}s_{\hat{\beta}_1}$, where $s_{\hat{\beta}_1} = \frac{s}{s_x \sqrt{n-1}}$

- Test for $H_0: \beta_1 = \beta_{10}$: $t = \frac{\hat{\beta}_1 - \beta_{10}}{s_{\hat{\beta}_1}} \sim t_{n-2}$ tests $H_0: \beta_1 = \beta_{10}$

Often we want to test $H_0: \beta_1 = 0$ vs. $H_A: \beta_1 \neq 0$. If $H_0$ is true, $y$ doesn’t __________ and regression ______________. Proceed with regression only if $H_0$ is __________.

e.g. Find a 95% confidence interval for $\beta_1$ and test $H_0: \beta_1 = 0$ for the sparrowhawk data.

Inference on the $y$-intercept, $\beta_0$

Experts say:

- Confidence interval for $\beta_0$: $\hat{\beta}_0 \pm t_{n-2,\alpha/2}s_{\hat{\beta}_0}$, where $s_{\hat{\beta}_0} = s \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_x^2}}$

- Test for $H_0: \beta_0 = \beta_{00}$: $t = \frac{\hat{\beta}_0 - \beta_{00}}{s_{\hat{\beta}_0}} \sim t_{n-2}$ tests $H_0: \beta_0 = \beta_{00}$

e.g. Find a 95% confidence interval for $\beta_0$. 
Inference on the mean response, $E(y) = \mu_y$, for a given $x$

Recall that our linear model is $y = \beta_0 + \beta_1 x + \varepsilon$. The mean response at $x$ is $E(y) = \mu = \mu_{\beta_0 + \beta_1 x + \varepsilon} = \mu_{\beta_0 + \beta_1 x}$

To estimate $\mu_y$, use $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ (draw).

Experts say:

- Confidence interval for $\mu_y = \beta_0 + \beta_1 x$: $[\hat{\beta}_0 + \hat{\beta}_1 x] \pm t_{n-2,\alpha/2} s_{\hat{y}}$, where $s_{\hat{y}} = s\sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{(n - 1)s_x^2}}$

- Test for $H_0: \mu_y = \beta_0 + \beta_1 x = \mu_0$: $\frac{\hat{y} - \mu_0}{s_{\hat{y}}} \sim t_{n-2}$ tests $H_0: \mu_y = \mu_0$

  e.g. Find a 95% confidence interval for $\mu_y = \beta_0 + \beta_1 x$ and test $H_0: \mu_y = 18$ for sparrowhawk data.

  e.g. Use R to check most of the inference work above.
**Extra Example**

e.g. 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>8.1</td>
<td>7.7</td>
<td>7.5</td>
<td>7.2</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 the correlation between additive concentration and drying time.

c. Find least-squares line.

d. Find fitted value and residual for each point.

e. If concentration is increased by .1%, by how much will drying time change?

f. Predict drying time for concentration = 4.4%.

g. For what concentration would you predict a drying time of 8.2 hours?

h. Test $H_0 : \beta_1 = 0$.

i. Find 95% confidence intervals for $\beta_0$ and $\beta_1$.

j. Find a 95% confidence interval for the mean drying time corresponding to an additive concentration of 4.9%.

e.g. Use R to check answers to the example, above (and to practice reading R output).