Regression (Lecture #17)

- **y** = tail length, **x** = wing length
- Regression line: slope = \( \frac{s_y}{s_x} \)
- Best line for predicting **y** from **x**?
- Gauss (~1800)
- Francis Galton (1890)

**Example:**

1. **n = 1050**
2. **ht father**
3. **ht sons**
4. **n = 1050**

**R - 25 (Eq. 17.21)**

- Slope of regression line:
  \[ b_1 = \frac{s_y}{s_x} \]
- Equation of regression line:
  \[ \hat{y} = \beta_0 + \beta_1 x \]
- Predicted **y** value
  \[ \beta_0 = \bar{y} - \beta_1 \bar{x} \]

**Care 2nd time**

**Care 1st time**

L-217: jump example
\[
\begin{bmatrix}
1 & 100 \\
2 & 94 \\
3 & 98 \\
50 & 5 \\
\end{bmatrix}
\]

Gauss

\[
\frac{1}{n} \sum_{i=1}^{n} [Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2
\]

- Find \((\hat{\beta}_0, \hat{\beta}_1)\) to minimize \(\frac{1}{n} \sum_{i=1}^{n} [Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2\)
- Result is least squares estimates

L-248 - Inference about slope

Inferential summary

<table>
<thead>
<tr>
<th>Unknown pop. quantity of main interest</th>
<th>(\hat{\beta}_1) = POP slope for predicting</th>
</tr>
</thead>
<tbody>
<tr>
<td>estimate</td>
<td>(\hat{\beta}_1 = 0.77 \text{ cmHg/cMWL})</td>
</tr>
<tr>
<td>S.E. (S.E.)</td>
<td>(\hat{\beta}_1 = 0.14 \text{ cmHg/cMWL})</td>
</tr>
<tr>
<td>95% CI</td>
<td>(\hat{\beta}<em>1 \pm t</em>{n-2,0.95} \text{ S.E.} (\hat{\beta}_1) = (0.40, 1.08))</td>
</tr>
</tbody>
</table>
FACTS

1. $E_{i i d} \left( \beta_i \right) = \beta_i$,

   "given"

2. $S_{E_{i i d}} \left( \beta_i \right) = S_{\hat{y} | x}$

   where

   $S_{\hat{y} | x} = \frac{S_y}{\sqrt{n-1}} \sqrt{1 - r^2} \cdot \sqrt{\frac{n-2}{n-2}}$

   residual = root mean squared error

   So

   $\delta_{\hat{y} | x} = \sqrt{\frac{1}{n-2} \sum_{i=1}^{n-2} e_i^2}$

   squared error

   is the regression practically useful?

   2 ways to answer:

   1. $r^2$, recall variance $V(y) = S_y^2 = \frac{1}{n-1} \sum_{i=1}^{n-1} (y_i - \bar{y})^2$

      in regression $y = \hat{y} + \epsilon$, so $V(y) = V(\hat{y} + \epsilon)$

      $V(\hat{y} + \epsilon) = V(\hat{y}) + V(\epsilon)$

      so $r^2 = \frac{V(\hat{y})}{V(y)} = \%$ of variance in $Y$ associated with regression of $Y$ on $X$

      - want $r^2$ to be large

   2. $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

      a. ignore $x$ or don’t know $x$

      predict $y$ anyway - best prediction is $\hat{y} = \bar{y}$

      give or take for $s_y$

      b. use $x$ to predict $y$, best estimate is

      $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$; give or take for this $\hat{y}$ is smaller
If scatterplot looked like this, simple linear regression not good.

This is a residual plot.

Good

Bad