we did regression (fitting) by assuming a model for data \((x_i, y_i)\):

\[ y_i = \alpha + \beta x_i + \varepsilon_i \quad \text{error/residual} \]

Obs. \(y\) at \(x_i\); \(y\) of line \(\hat{y}(x) = \alpha + \beta x\) at \(x = x_i\).

To find the "best" \(\alpha, \beta\) (i.e., line), we minimized \(\text{SSE}\):

\[
\text{SSE} = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left[ y_i - (\alpha + \beta x_i) \right]^2
\]

\(\text{Obs} \rightarrow \text{pred.} \quad \text{Comp.}! \)

and got \(\hat{\beta} = \frac{\sum x_i y_i - \bar{x} \sum y_i}{\sum x_i^2 - \bar{x}^2} \quad \hat{\alpha} = \bar{y} - \hat{\beta} \bar{x} \).

Then, the eqn of the "best fit" is \(\hat{y}(x) = \hat{\alpha} + \hat{\beta} x\).

Note: \(\hat{y}(x_i) = \hat{\alpha} + \hat{\beta} x_i\) (no \(\varepsilon\) !)

\( \Rightarrow \hat{y}(x_i) \) is sometimes written as \(\hat{y}_i\).

Regression is most useful when \(x = \) Easy, \(y = \) Hard to measure.

E.g. \(x = \) Blood flow velocity (FV) with ultrasound, \(y = \) Intracranial Pressure (ICP).

When \(\hat{\alpha}, \hat{\beta}\) are obtained from regression, then, given \(x\), we can predict \(y\) from \(\hat{y}(x) = \hat{\alpha} + \hat{\beta} x\). (No \(\varepsilon_i\) !)

Also \(\hat{\beta} = \frac{S_{xy}}{S_{xx}} \quad \text{where} \quad S_{xx} = \sum (x_i - \bar{x})^2 \quad S_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y})\).
Example

<table>
<thead>
<tr>
<th>height (x)</th>
<th>weight (y)</th>
<th>xy</th>
<th>x²</th>
</tr>
</thead>
<tbody>
<tr>
<td>72</td>
<td>200</td>
<td></td>
<td></td>
</tr>
<tr>
<td>65</td>
<td>180</td>
<td></td>
<td></td>
</tr>
<tr>
<td>68</td>
<td>120</td>
<td></td>
<td></td>
</tr>
<tr>
<td>70</td>
<td>150</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[
\bar{x} = \frac{\sum x}{n} = \frac{65 + 68 + 70}{3} = 68\frac{2}{3} \\
\bar{y} = \frac{\sum y}{n} = \frac{200 + 180 + 120 + 150}{4} = 170 \\
\bar{xy} = \frac{\sum xy}{n} = \frac{72 \times 200 + 65 \times 180 + 68 \times 120 + 70 \times 150}{4} = 10020 \\
\bar{x^2} = \frac{\sum x^2}{n} = \frac{65^2 + 68^2 + 70^2}{3} = 10276.67 \\
\]

\[
\hat{b} = \frac{\bar{xy} - \bar{x} \bar{y}}{\bar{x^2} - \bar{x}^2} = \frac{10020 - 68 \frac{2}{3} \times 170}{10276.67 - 68 \frac{2}{3}^2} = \frac{11224.8 - 69(161.6)}{4766.6 - 69(69)} = 13.28 \\
\]

\[
\hat{a} = \bar{y} - \hat{b} \bar{x} = 161.6 - 13.28 \times 69 = -755.11 \\
\]

\[
\text{Interpret: A change of 1 in } x \text{ is associated with an avg. change of 13.28 pounds.} \\
\]

\[
\Rightarrow \hat{\gamma}(x) = -755.11 + 13.28x \\
\]

\[
\Rightarrow \text{E.g. Joe's predicted weight, according to his height, is } \hat{\gamma} = 13.28(70) - 755.11 = 174.9 \text{ pounds.} \\
\]

\[
\Rightarrow \text{We can now predict everyone's weight, from their height.} \\
\]

\[
\text{Height (x)} \quad \text{Weight (y)} \quad \text{(y - \hat{y})} \\
\hline
72 \quad 200 \quad \ldots \quad 201.5 \quad -1.5 \\
65 \quad 180 \quad 174.9 \quad 5.1 \\
68 \quad 120 \quad 108.5 \quad 11.5 \\
70 \quad 150 \quad 148.3 \quad -30.3 \\
\]

\[
\text{For the people in the data set, we can also find their error/residual.} \\
\]

\[
\Rightarrow \text{For people outside the data set (e.g., Jane) we can predict their y from their x, but we cannot compute error, because we don't know their true y. In Ch 9, we'll address this issue.} \\
\]

\[
\text{However, be WARNED if you extrapolate} \Rightarrow x=0 \implies y=-755 \text{ pounds!} \\
\]

Q1: For a given data set on \((x,y)\), the OLS fit to \(y \text{ vs. } x\) is \(\hat{a} + \hat{b}x\). For the same data, the slope of the OLS fit to \(x \text{ vs. } y\) is

\[ a) \hat{b} \quad b) \frac{1}{\hat{b}} \quad c) 1 \quad d) \text{None of the above}. \]

I.e. what does switching \(x \leftrightarrow y\) do to \(\hat{b}\)?

\[ \hat{b} = \frac{\bar{x} \bar{y} - \bar{xy}}{\bar{x}^2 - \bar{x}} \rightarrow \frac{\bar{x} - \bar{y}}{\bar{y}^2 - \bar{y}} = \frac{1}{\hat{b}} \]

Shifting gears again.

There is a different (more useful) way of looking at regression, via variance. This way, we will arrive at quantities called \(R^2\) and \(s_e\) which together assess how good the fit is.

Let me motivate it:

\( \rightarrow \) Suppose we measure a table's length.

\( \rightarrow \) Repeat, and histogram:

\[ s_y = \sqrt{\frac{S_{yy}}{n-1}} \approx 10 \]

\[ \text{True length } = 150 \pm 10 \text{ cm} \]

\( \rightarrow \) Now, suppose you are unhappy with the large \(s_y\).

\( \rightarrow \) You may wonder, could some of that variability be due to something else that is varying every time you make a measurement of \(y\)? \(\text{X} = \text{temperature? humidity?}\)

If so, then by measuring \(y\) and \(x\), we may be able to reduce the \(\pm\) of our report, by specifying \(y\) at a given \(x\).
Analysis of Variance (ANOVA) approach to regression:

How much of the variation in $y$ is due to the (linear) relationship between $y$ and $x$?

Variance of $y = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2$

\[ SS_{\text{total}} = SS_{\text{explained}} + SS_{\text{unexplained}} \]

- Total variation in $y$.
- Variation in $y$ explained by (or due to) $x$.
- Variation in $y$ unexplained by $x$.

\[ SST = SS_{\text{explained}} + SSE \]

Variability is reduced from $\pm \sigma^2$ to something smaller, say $\sigma_1^2$.

Therefore, $\frac{SS_{\text{explained}}}{SST} \times 100 \%$ called $R^2$ measures how good the fit is.

$0 \leq R^2 \leq 1$ (Good model/fit)

The other piece, $SS_{\text{unexplained}} = SSE$, is a sum (of squares), and so can be can be “averaged” to provide a measure of typical error.

\[ \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{1}{n-2} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2} = Se \sim \text{std. dev. of errors} \]

“Fuzzy Avg.” error $\sim$ typical error. Compare with:

\[ S_y^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2 \]

Report $\hat{y}(x) \pm Se$.
Example (same as in last few lectures):

\[ \text{SS} = \sum \frac{1}{n} (y_i - \overline{y})^2 = \ldots = 6251.2 \]

\[ \text{SS}^2 = \sum (y_i - \overline{y})^2 = \text{last column in table in prev. lecture.} \]

\[ = (-1.5)^2 + (5.1)^2 + (11.5)^2 + (-30.3)^2 + (15.1)^2 = 1307 \]

\[ R^2 \text{ Coef. of det.} = \frac{\text{SST} - \text{SSE}}{\text{SST}} = \frac{6251.2 - 1307}{6251.2} = 0.79. \]

Conclusion: 79\% of the variability (or variation) in \( y \) (weight, table's length) is due to (can be explained by) the linear relation with \( x \) (height, or temperature).

The other piece of the decomposition:

\[ s_e = \sqrt{\frac{1307}{5 - 2}} = 20.9 \text{ pounds} \]

Conclusion: The typical deviation of the \( y \) values (weight/Table's length) (i.e. error or residual) about the fit is about 21 pounds.

Report weight (or Table length): \( \hat{y} \pm 20.9 \) with \( R^2 = 0.79 \)

or ICP

or \[ -755 + 13.3 \times \text{height or DV or } \ldots \]
For the data shown here:
\[ x = 45, 58, 71, 71, 85, 98, 108 \]
\[ y = 3.20, 3.40, 3.47, 3.55, 3.60, 3.70, 3.80 \]
a) Compute the eq. of the OLS fit.
b) Compute the total variation, SST.
c) Decompose it into explained and unexplained.
d) Compute R2 and interpret it (in English),
e) Compute the std. dev of errors, s_e, and interpret it (in English).

All by hand. You may use R to compute sums, means, std. deviations, but not a function that does regression or analysis of variance.

Consider the following decomposition:
\[
\sum_i (y_i - \bar{y})^2 = \sum_i [(\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)]^2
\]
\[
= \sum_i (\hat{y}_i - \bar{y})^2 + \sum_i (y_i - \hat{y}_i)^2 + 2 \sum_i (\hat{y}_i - \bar{y})(y_i - \hat{y}_i).
\]

In past hws I have asked students to prove that the last term is zero if \( \hat{y}_i = \hat{\beta} x_i + \hat{\alpha} \), with \( \hat{\alpha}, \hat{\beta} \) being the OLS estimates (i.e. \( \hat{\beta}, \hat{\alpha} \) given in lectures, book). Unfortunately, it's a long calculation; so this time we'll try to show that it's zero using simulation in R. Write code to:

a) generate a sample of size 100 from the uniform dist. between -1 and +1. Call it \( x \).
b) generate \( y \) such that \( y = 2 + 3 x_i + \epsilon_i \) with \( \epsilon_i \) having a normal dist. with \( \mu = 0, \sigma = 0.5 \).
c) Do regression on \( x, y \), and call the predictions \( \hat{y} \).
d) compute \( \sum (\hat{y}_i - \bar{y})(y_i - \hat{y}_i) \). It should be (very) zero!
SS_{exp} can be computed from its defining relation: \( \sum_i (\hat{y}_i - \bar{y})^2 \)

Or from \((SS_T - SSE)\), or from \(\hat{\beta}\) and \(S_{xx}\), as follows.

Explain what has happened at every step.

\[
SS_{exp} = \sum_i (\hat{y}_i - \bar{y})^2 \\
= \sum_i (\hat{\alpha} + \hat{\beta} x_i - \bar{y})^2 \\
= \sum_i (\bar{y} - \hat{\beta} \bar{x} + \hat{\beta} x_i - \bar{y})^2 \\
= \sum_i (\hat{\beta})^2 (x_i - \bar{x})^2 \\
= (\hat{\beta})^2 \sum_i (x_i - \bar{x})^2 \\
= (\hat{\beta})^2 S_{xx}
\]

If you would like, print out this page, and write your answers in the space here.