Lecture 12 (C2.3)  

"curve fitting"

Basic picture of regression:

\[ y = \alpha + \beta x \]

\[ \hat{y}_3 = \hat{\alpha} + \hat{\beta} x_3 \]

\[ (x_3, \alpha + \beta x_3) \]

\[ (x_3, y_3) \]

\[ \text{error} \]

\[ \hat{\beta} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \]

\[ \hat{\alpha} = \bar{y} - \hat{\beta} \bar{x} \]

\[ s = \sqrt{n - 2} \cdot \frac{\sum (y_i - \hat{y}_i)^2}{n-2} \]

Think of \( x \): Blood Flow Velocity (FV) measured non-invasively.

\( y \): Intracranial Pressure (ICP) measured invasively.

The book introduces the notation

\[ S_{xx} = \sum (x_i - \bar{x})^2 \]

\[ S_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}) \]

Numerators of sample var. \( s^2 \), \( s_y^2 \).

in which case it's easy to show that

\[ \hat{\beta} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}} \]

Finally, note the resemblance between the formulas for \( \hat{\beta} \) and \( r \). But their meaning is completely different.

\[ \frac{S_{xy}}{\sqrt{S_{xx} \cdot S_{yy}}} \]
Example

<table>
<thead>
<tr>
<th>Height (x)</th>
<th>Weight (y)</th>
<th>xy</th>
<th>x²</th>
</tr>
</thead>
<tbody>
<tr>
<td>65</td>
<td>120</td>
<td>7800</td>
<td>4225</td>
</tr>
<tr>
<td>68</td>
<td>118</td>
<td>7864</td>
<td>4624</td>
</tr>
<tr>
<td>70</td>
<td>190</td>
<td>13300</td>
<td>4900</td>
</tr>
</tbody>
</table>

\[
\bar{x} = \frac{65 + 68 + 70}{3} = \frac{193}{3} = 64.3333
\]

\[
\bar{y} = \frac{120 + 118 + 190}{3} = \frac{428}{3} = 142.6667
\]

\[
\hat{b} = \frac{\sum xy - \bar{y} \cdot \sum x}{\sum x^2 - (\sum x)^2} = \frac{11224.8 - 69(161.6)}{4766.6 - 67(69)} = 13.28
\]

\[
\hat{a} = \bar{y} - \hat{b} \bar{x} = 161.6 - 13.28(69) = -755
\]

\[\hat{y} = \hat{a} + \hat{b}x\]

\[\hat{y} = -755 + 13.28x\]

Interpret: A change of 1 in x is associated with an avg. change of 13.28 pounds.

For example, Joe's predicted weight, according to his height, is

\[\hat{y} = 13.28(70) - 755.11 = 174.9\] pounds.

We can now predict everyone's weight:

<table>
<thead>
<tr>
<th>Height (x)</th>
<th>Weight (y)</th>
<th>Predicted y</th>
<th>Error (or residuals)</th>
</tr>
</thead>
<tbody>
<tr>
<td>72</td>
<td>200</td>
<td>201.5</td>
<td>-1.5</td>
</tr>
<tr>
<td>Joe</td>
<td>70</td>
<td>180</td>
<td>174.9</td>
</tr>
<tr>
<td>65</td>
<td>120</td>
<td>108.5</td>
<td>148.3</td>
</tr>
<tr>
<td>68</td>
<td>118</td>
<td>148.3</td>
<td>174.9</td>
</tr>
</tbody>
</table>

even if they are not in the data set. Need x, though.

However, be WARNED if you extrapolate; see next page for another example of bad extrapolation.

Q1: For the above example, when \( \hat{y} = -755 + 13.28x \), the SSE is given by \((-1.5)^2 + (5.1)^2 + \ldots + (15.1)^2\). The SSE of any other line (\( \hat{y} = \ldots \)) will be

A) greater
B) smaller
C) equal to
D) will depend.

We minimized SSE to get \( \hat{y} = -755 + 13.28x \).
What is most interesting in this relationship is that it can be extrapolated back to the origin of life. Genome complexity reaches zero, which corresponds to just one base pair, at time ca. 9.7 billion years ago (Fig. 1). A sensitivity analysis gives a range for the extrapolation of ±2.5 billion years (Sharov, 2006). Because the age of Earth is only 4.5 billion years, life could not have originated on Earth even in the most favorable scenario (Fig. 2). Another complexity measure yielded an estimate for the origin of life date about 5 to 6 billion years ago, which is similarly not compatible with the origin of life on Earth (Jørgensen, 2007). Can we take these estimates as an approximate age of life in the universe? Answering this question is not easy because several other problems have to be addressed. First, why the increase of genome complexity follows an exponential law instead of fluctuating erratically? Second, is it reasonable to expect that biological evolution had started from something equivalent in complexity to one nucleotide? And third, if life is older than the Earth and the Solar System, then how can organisms survive interstellar or even intergalactic transfer? These problems as well as consequences of the exponential increase of genome complexity are discussed below.

"Life before Earth"

Confidence Bands (Ch. 11)

Life predates Earth!

Extrapolation can lead to dangerous conclusions!

Actually NO!
There is a different (more useful) way of looking at this via variance. This way, we will arrive at quantities called $R^2$ and $S_e$, which together assess how good the fit is.

- Suppose we measure table length $y$.
- Repeat, and histogram:
  
  \[ s_y = \sqrt{\frac{\sum y^2}{n-1}} \approx 0 \]
  
  True length = 150 ± 10 cm

- Now, suppose you are unhappy with the large $s_y$.
- You may wonder, could some of that variability be due to something else that is varying every time you make a measurement of $y$? $x =$ 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 $x$.

Q2: Suppose we have $n$ observations of $y$ ($y_i$, $i=1, \ldots, n$), but we have no data on $x$, and we want to predict some “future” (e.g. next) $y$-value. What number should you use for your prediction?

Hint: I have not taught you the answer! Just think!

(A) $\bar{y}$  (B) min($y$)  (C) max($y$)  (D) The “last” $y$

$\bar{y}$ is the typical $y$.

BTW: you would report Table length: $\bar{y} \pm s_y$. 
Analysis of Variance (ANOVA):

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

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

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

\[ S_{y-x} = \sum_{i} (y_i - \hat{y}_i)^2 \]

\[ S_{x} = \sum_{i} (\hat{y}_i - \bar{y})^2 \]

Total variation in $y$.

Variation in $y$ explained by (or due to) $x$.

Variation in $y$ unexplained by $x$.

\[ S_{total} = S_{explained} + S_{unexplained} \]

\[ SST = S_{explained} + SSE \]

Variability is reduced from $\pm (10)^2$ to something smaller, say $\pm (3)^2$.

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

Percent variation in $y$, explained by $x$.

\[ (Bad \ Model/\ fit) \ 0 \leq R^2 \leq 1 \ (Good \ Model/\ fit) \]

The other piece, $SS_{unexplained} = SSE$, is a sum-squared, and so can be “Averaged” to provide a measure of typical error. Specifically,

\[ \sqrt{\frac{SSE}{n-2}} = \hat{e} \]

Typo in book, p. 121

\[ \hat{e} \approx \text{std. dev. of errors} \sim \text{typical error}, \]
When there is a linear relationship between $\mathbf{x}$ and $\mathbf{y}$, then some portion of the variation in $\mathbf{y}$ can be attributed to (or explained by) $\mathbf{x}$. That portion is $\text{SS}_{\text{expl.}}$, and the (unexplained) rest is $\text{SS}_{\text{unexp}} = \text{SSE}$.

So the variability in $\mathbf{y}$, $\text{SST}$, is reduced to $\text{SSE}$.
Example (from previous lecture):

\[ \text{SST} = \sum_i (y_i - \bar{y})^2 = \ldots = 6251.2 \]

\[ \text{SSE} = \sum_i (y_i - \hat{y}_i)^2 = \text{last column in table in prev. lecture} = (-1.5)^2 + (5.1)^2 + (11.5)^2 + (-30.3)^2 + (45.1)^2 = 1307 \]

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

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

The other piece of the decomposition:

\[ s_e = \sqrt{\frac{\text{SSE}}{n-2}} = 20.9 \text{ pounds} \]

Conclusion: The typical deviation of the data about the fit (i.e. error) is about 21 pounds.

Report Table length: \( \hat{y} = 20.9 - 755 + 13.3 \times \)

\[ R^2 = 0.79 \]
Come up with another example of $x$ and $y$ (like FV and ICP), where regression can help in predicting $y$ from $x$ in a situation where without regression the “cost” of measuring $y$ directly is extremely high (like ICP).

Suppose all we have are data on a single variable $y$: $y_i, i = 1, 2, 3, \ldots, n$. Show that the predictor that minimizes SSE is the sample mean $\bar{y}$. Hint: let $\hat{y}$ denote the prediction, and then minimize SSE.

Consider the following decomposition:

$$\sum_i (y_i - \bar{y})^2 = \sum_i [(\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)]^2$$

$$= \sum_i (y_i - \bar{y})^2 + \sum_i (y_i - \hat{y}_i)^2 + 2 \sum_i (y_i - \bar{y})(y_i - \hat{y}_i)$$

In past hws I have asked students to prove that the last term is zero if $\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$, with $\hat{\alpha}, \hat{\beta}$ being the OLS estimates (i.e. $\hat{\alpha}, \hat{\beta}$ given in lecture, 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 distribution between -1 and 1. Call it $x$.

b) Generate $y$ such that $y = 2 + 3x + \epsilon$ with $\epsilon$ having a normal distribution with $\mu = 0, \sigma = 0.5$.

c) Do regression on $x, y$, and call the predictions $\hat{y}$.

d) Compute $\sum_i (\hat{y}_i - \bar{y})(y_i - \hat{y}_i)$. It should be (very) zero!
For the data shown in problem 3.22
a) Compute the eqn of the OLS fit
b) Compute the total variation, SST.
c) Decompose it into explained and unexplained.
d) Compute $R^2$, and interpret (in English).
e) Compute the std. dev. of errors, and interpret (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.

Ignore!

Select 10 random numbers between 0 and 100 (from any distr.)