1. KNNL Problem 1.3
If, for a given elapsed time since the molding process \((X)\), there is no chance variation in the hardness \((Y)\), then the relationship is mathematical. In this case, however, there could be variability in the state of the plastic at the termination of the molding process and in the measurement of the hardness. Both these uncertainties would justify more of a regression approach. What is unclear from the description is whether the relationship is expected to be linear.

2. KNNL Problem 1.5
No. The mean of \(Y\) is a constant. By including the \(\varepsilon_i\) term, the student is saying the mean is a random variable. If this term is dropped, the statement would be correct.

3. KNNL Problem 1.22
(a) It can be shown using the SAS code at the end of this answer key that the estimated regression line is \(\hat{Y} = 168.6 + 2.034X\). The plot of \(Y\) vs \(X\) is reasonably linear with a roughly even number of observations above and below the fitted line (not shown).
(b) For \(X = 40\) hours, \(\hat{Y} = 249.975\).
(c) The change in \(Y\) for an increase of 1 hour is simply the slope (2.034).

4. KNNL Problem 1.26
(a) It is easy to see that the residuals sum to zero.
(b) \(\sigma^2\) is estimated by MSE = 10.4589, and \(\sigma\) is estimated by “Root MSE” = 3.2340. \(\sigma\) is expressed in Brinell units.

5. A regression analysis of a set of data produced the following fitted equation: \(\hat{y} = 3 + 8x\).
(a) If \(x\) increases 5 units, how does \(\hat{y}\) change?
The slope describes the rate of change in \(Y\) so a 5 unit increase in \(X\) results in a \(5 \times 8 = 40\) unit change in the fitted value.
(b) Here \(x\) is measured in degrees Celsius. Rewrite the model with \(x\) replaced by \(x^*\) where \(x^*\) is \(x\) expressed in degrees Fahrenheit. Use the fact that \(x = (5/9)(x^* - 32)\).
Since \(x = 0\) is comparable to \(x^* = 32\), the equations should result in the same value for \(y\). This is a helpful check. Given that \(x = (5/9)(x^* - 32)\), we rewrite the equation as \(y = 3 + 8((5/9)(x^* - 32) + \varepsilon = -1257/9 + (40/9)x^* + \varepsilon\).

6. KNNL Problem 1.39
a. It is subject to show identical \(b_1\) from both data. Consider a more general case with observations \((X_i, Y_i), i = 1, \cdots, 6\), but \(X_1 = X_2, X_3 = X_4,\) and \(X_5 = X_6\). So, we have
averaged observations \((\bar{X}_i, \bar{Y}_i), i = 1, 2, 3\) with \(\bar{X}_i = X_{2i-1}\) and \(\bar{Y}_i = (Y_{2i-1} + Y_{2i})/2\). Then
\[
\sum_{i=1}^{6} (X_i - \bar{X})^2 = 2 \sum_{i=1}^{3} (\bar{X}_i - \bar{X})^2
\]
and
\[
\sum_{i=1}^{6} (X_i - \bar{X})(Y_i - \bar{Y}) = \sum_{i=1}^{3} (\bar{X}_i - \bar{X})(Y_{2i-1} - \bar{Y} + Y_{2i} - \bar{Y})
= \sum_{i=1}^{3} (\bar{X}_i - \bar{X})(2\bar{Y}_i - 2\bar{Y})
= 2 \sum_{i=1}^{3} (\bar{X}_i - \bar{X})(\bar{Y}_i - \bar{Y})
\]
Therefore, we have identical estimate \(b_1\) using both datasets.

b. Let us consider a special case with observations \((X, Y_i), i = 1, \cdots, n\) (note observing constant \(X\) in all observations). Then an unbiased estimate of the error term variance is
\[
\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}{(n - 1)}.
\]
For the data \((\bar{X}_i, Y_{2i-1})\) and \((\bar{X}_i, Y_{2i})\), we have the unbiased estimate
\[
\hat{\sigma}^2_i = \{(Y_{2i-1} - \bar{Y}_i)^2 + (Y_{2i} - \bar{Y}_i)^2\}/(2 - 1).
\]
Therefore, we have the estimate of the error term variance
\[
\hat{\sigma}^2 = \frac{\hat{\sigma}^2_1 + \hat{\sigma}^2_2 + \hat{\sigma}^2_3}{3}.
\]

7. KNNL Problem 1.40
An observation that ends up falling directly on the fitted line does not contribute at all to the least squares quantity \(Q\). This means that the intercept and slope estimates do not depend on this observation and so removing it will not affect the fitted line. Removing the observation, however, does reduce the sample size to \(n - 1\). This changes the degrees of freedom, so our error variance will be different.

8. KNNL Problem 1.41
(a) Since
\[
Q = \sum_{i=1}^{n} (Y_i - \beta_1 X_i)^2 = \sum_{i=1}^{n} Y_i^2 - 2\beta_1 \sum_{i=1}^{n} X_i Y_i + \beta_1^2 \sum_{i=1}^{2} X_i^2
\]
Let \(\partial Q/\partial \beta_1 = 0\), we have the least squares estimator of \(\beta_1\) as follows,
\[
b_1 = \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2}.
\]
(b) The likelihood function for the linear model is

\[ L(\beta_1) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - \beta_1 X_i)^2 \right] \]

taking logarithm on both sides of

\[ \mathcal{L} = \log L(\beta_1) = -\frac{n}{2} \log 2\pi - \frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - \beta_1 X_i)^2 \]

and taking derivative with respect to \( \beta_1 \):

\[ \frac{d\mathcal{L}}{d\beta_1} = \frac{1}{\sigma^2} \sum_{i=1}^{n} X_i (Y_i - \beta_1 X_i) \]  \hspace{1cm} (1)

Now we set the equation to zero, and get:

\[ \hat{\beta}_1 = \frac{\sum X_i Y_i}{\sum X_i^2} \]

It is exactly the same as the least square estimator.

(c) The estimator is unbiased following the fact

\[ E[b_1] = \sum_{i=1}^{n} X_i E[Y_i] / \sum_{i=1}^{n} X_i^2 = \sum_{i=1}^{n} X_i^2 \beta_1 / \sum_{i=1}^{n} X_i^2 = \beta_1. \]

/* SAS Codes for KNNL 1.22 & 1.26 */
DATA hardness;
  INPUT y x @@;
CARDS;
  199.0 16.0 205.0 16.0 196.0 16.0 200.0 16.0
  218.0 24.0 220.0 24.0 215.0 24.0 223.0 24.0
  237.0 32.0 234.0 32.0 235.0 32.0 230.0 32.0
  250.0 40.0 248.0 40.0 253.0 40.0 246.0 40.0
;

SYMBOL1 V=CIRCLE I=RL;
PROC GPLOT;
  PLOT y*x;
RUN;

PROC REG;
  MODEL y=x /P;
RUN;
QUIT;