1. REGRESSION THROUGH THE ORIGIN

- Model: \( Y_i = \beta_1 X_i + \epsilon_i \) for \( i = 1,\ldots,n \)

  where \( \beta_1 \) is a parameter (unknown constant),
  \( X_i \) are known constants, and
  \( \epsilon_i \) are independent \( N(0,\sigma^2) \).

- Regression function: \( E\{Y\} = \beta_1 X \)

- Least squares estimate of \( \beta_1 \): \( b_1 = \frac{\sum X_i Y_i}{\sum X_i^2} \)

- Fitted values: \( \hat{Y}_i = b_1 X_i \)

- \((1 - \alpha)\) confidence limits for \( \beta_1 \): \( b_1 \pm t_{(1-\alpha/2, n-1)} \sqrt{\frac{\text{MSE}}{\sum X_i^2}} \)

- \((1 - \alpha)\) confidence limits for \( E\{Y_h\} \), the mean response when \( X = X_h \):

  \[
  b_1 X_h \pm t_{(1-\alpha/2, n-1)} \sqrt{\frac{\text{MSE} \cdot X_h^2}{\sum X_i^2}}
  \]

- \((1 - \alpha)\) prediction limits for new response \( Y_{h\text{ (new)}} \):

  \[
  b_1 X_h \pm t_{(1-\alpha/2, n-1)} \sqrt{\text{MSE} \left( 1 + \frac{X_h^2}{\sum X_i^2} \right)}
  \]

- For regression through the origin, MINITAB uses \( \text{SSTO} = \sum Y_i^2 \), \( \text{SSE} = \sum (Y_i - \hat{Y}_i)^2 \),
  and \( \text{SSR} = \sum Y_i \hat{Y}_i \). Here SSR could be negative. Hence, for regression through the origin,
  the coefficient of determination, \( r^2 \), has no clear meaning.

- Note that for regression through the origin the sum of residuals, \( \sum \epsilon_i \), does not usually equal 0.

- Note that it is usually better not to use regression through the origin.
**Example**: (Chapter 4, problem 12). Data

<table>
<thead>
<tr>
<th>$X_i$</th>
<th>7</th>
<th>12</th>
<th>10</th>
<th>10</th>
<th>14</th>
<th>25</th>
<th>30</th>
<th>25</th>
<th>18</th>
<th>10</th>
<th>4</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Y_i$</td>
<td>128</td>
<td>213</td>
<td>191</td>
<td>178</td>
<td>250</td>
<td>446</td>
<td>540</td>
<td>457</td>
<td>324</td>
<td>177</td>
<td>75</td>
<td>107</td>
</tr>
</tbody>
</table>

where $X =$ number of galleys for a manuscript, and $Y =$ cost of correcting typos.

We obtain $\sum X_i^2 = 3215$, $\sum X_i Y_i = 57961$, and $\text{SSE} = 223.424$.

Here $b_1 = \frac{\sum X_i Y_i}{\sum X_i^2} = 57961/3215 = 18.0283$.

So the estimated regression function is $\hat{Y} = 18.0283 X$.

Find a 95% P.I. for a new response when $X = 10$. $\hat{Y}_{\text{new}} = 18.0283(10) = 180.283$.

$t_{0.975, 11} = 2.201$, $\sqrt{\frac{\text{MSE}}{1 + \frac{X_i^2}{\sum X_i^2}}} = \sqrt{\frac{223.424}{11} \left(1 + \frac{10^2}{3215}\right)} = 4.5763565$

Thus we have $180.283 \pm (2.201)(4.5763565) \Rightarrow 180.283 \pm 10.073 \Rightarrow (170.21, 190.36)$

**BONFERRONI INEQUALITY**

Let $\{A_i\}$ represent a set of $g$ events. Let $\overline{A}_i$ denote the complement of $A_i$.

Then $P(\overline{A}_1 \cap \overline{A}_2 \cap \ldots \cap \overline{A}_g) = 1 - P(\cup A_i)$

$\geq 1 - P(A_1) - P(A_2) - \ldots - P(A_g)$.

In other words, $P(\cup \overline{A}_i) \geq 1 - \sum P(A_i)$ since $P(\cup A_i) \leq \sum P(A_i)$.

Bonferroni Inequality: 

$$P(\cap \overline{A}_i) \geq 1 - \sum P(A_i)$$

If $P(A_i) = \alpha$ for all $i$, then $P(\cap \overline{A}_i) \geq 1 - g\alpha$.

For example, if $g = 2$ and $P(A_1) = P(A_2) = \alpha$, then $P(\overline{A}_1 \cap \overline{A}_2) \geq 1 - 2\alpha$.

If one wanted confidence intervals for $\beta_0$ and $\beta_1$ that had a family confidence coefficient of at least $1 - \alpha$, then one would construct confidence intervals that had individual confidence level of $1 - \alpha/2$. Hence, we would use

$$b_0 \pm t_{1-\alpha/4, n-2} \cdot s\{b_0\} \text{ for } \beta_0 \text{ confidence interval}$$

and

$$b_1 \pm t_{1-\alpha/4, n-2} \cdot s\{b_1\} \text{ for } \beta_1 \text{ confidence interval}$$

For example, if a family confidence coefficient of 0.95 is desired, use $t_{0.9875, n-2}$ since

$1 - \alpha = .95$ implies that $\alpha = .05$ and so $1 - \alpha/4 = 1 - .05/4 = 1 - .0125 = .9875$. 


In a multiple regression problem with model \( Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i \), if one wanted confidence intervals for \( \beta_0, \beta_1, \) and \( \beta_2 \) that had a family confidence coefficient of at least \( 1 - \alpha \), then one would construct confidence intervals that had individual confidence level of \( 1 - \alpha/3 \). Hence, we would use

\[
\begin{align*}
    b_0 &\pm t_{(1-\alpha/6, n-3)} s\{b_0\} \quad \text{for } \beta_0 \text{ confidence interval} \\
    b_1 &\pm t_{(1-\alpha/6, n-3)} s\{b_1\} \quad \text{for } \beta_1 \text{ confidence interval} \\
    b_2 &\pm t_{(1-\alpha/6, n-3)} s\{b_2\} \quad \text{for } \beta_2 \text{ confidence interval}
\end{align*}
\]

In this case, if a family confidence coefficient of .95 is desired, use \( t_{.90167, n-3} \) since \( 1 - \alpha = .95 \) implies that \( \alpha = .05 \) and so \( 1 - \alpha/4 = 1 - .05/4 = 1 - .0125 = .9875 \).

- The Bonferroni approach can be used for \textit{joint confidence intervals for mean responses}. When \( E\{Y_h\} \) is to be estimated for \( g \) levels \( X_h \), with family confidence coefficient \( 1 - \alpha \), the \textbf{Bonferroni confidence limits} are

\[
\hat{Y}_h \pm t\left(1 - \frac{\alpha}{2g}, n-2\right) s\{\hat{Y}\}.
\]

where \( s\{\hat{Y}\} = \sqrt{\text{MSE} \left( \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\text{SSX}} \right)} \)

*The Working-Hotelling limits use \( \sqrt{2F(1 - \alpha, 2, n-2)} \) in place of the the t-value. (See text.)

- The Bonferroni approach can be used for \textit{joint prediction intervals for new observations}. The \( 1 - \alpha \) simultaneous prediction limits are

\[
\hat{Y}_h \pm t\left(1 - \frac{\alpha}{2g}, n-2\right) s\{Y_{h(new)}\}.
\]

where \( s\{Y_{h(new)}\} = \sqrt{\text{MSE} \left( 1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\text{SSX}} \right)} \).

*The Scheffe' procedure uses \( \sqrt{gF(1 - \alpha, g, n-2)} \) in place of the t-value.

- Read examples and comments in text. Work problems.
INVERSE PREDICTIONS

If one wants to predict the value of X which gave rise to a new observation Y, one uses inverse prediction. Let $Y_{h(new)}$ represent the new Y value and let $X_{h(new)}$ represent the X value to be estimated. A point estimate of $X_{h(new)}$ is given by

$$\hat{X}_{h(new)} = \frac{Y_{h(new)} - b_0}{b_1}.$$

The approximate $(1-\alpha)$ confidence limits for an interval estimate of the unknown $X_{h(new)}$ are

$$\hat{X}_{h(new)} \pm t_{(1-\alpha/2, n-2)} \sqrt{\frac{MSE}{b_1^2} \left( 1 + \frac{1}{n} + \frac{(\hat{X}_{h(new)} - \bar{X})^2}{SSX} \right)}.$$

Example. Suppose $n = 20$, $\bar{X} = 25$, $SSX = 250$, $SSE = 40$ and $\hat{Y} = 10 + 2X$.

Find a 95% confidence interval for X when $Y = 50$.

$$\hat{X}_{h(new)} = \frac{50 - 10}{2} = 20; \quad t_{0.05, 18} = 2.101; \quad s_{\text{predX}} = \sqrt{\frac{40/18 \left( 1 + \frac{1}{20} + \frac{(20-25)^2}{250} \right)}{4}} = 0.799$$

So we obtain $20 \pm (2.101)(0.799) \Rightarrow 20 \pm 1.68 \Rightarrow (18.32, 21.68)$

CHOICE OF X LEVELS

- Choice of X levels depends on objective of regression analysis.

<table>
<thead>
<tr>
<th>Objective</th>
<th>X levels</th>
</tr>
</thead>
<tbody>
<tr>
<td>estimate $\beta_1$</td>
<td>place observations at extremes (linearity assumed)</td>
</tr>
<tr>
<td>estimate $\beta_0$</td>
<td>have $\bar{X} = 0$</td>
</tr>
<tr>
<td>estimate 1 mean resp. or predict new</td>
<td>have $\bar{X} = X_h$</td>
</tr>
<tr>
<td>estimate several mean responses</td>
<td>have $\bar{X}$ in the center of the $X_h$ levels of interest</td>
</tr>
</tbody>
</table>

D.R. Cox’s suggestions:

(i) only 2 levels are needed to check for effect and its direction
(ii) 3 levels needed whenever slope and curvature are desired
(iii) 4 levels for further examination of response curve shape
(iv) more levels for finding min or max response or finding more detailed shape.

Generally satisfactory (except in last case) to use equal-spacing and equal number of obs.

Read text.