Regression through the origin

 Sometimes due to the nature of the problem (e.g. (i) physical law where one variable is proportional to another variable, and the goal is to determine the constant of proportionality; (ii) \( X = \text{sales}, Y = \text{profit from sales} \)), or, due to empirical considerations ( in the full regression model the intercept \( \beta_0 \) turns out to be insignificant), one may fit the model \( Y_i = \beta_1 X_i + \epsilon_i \), where \( \epsilon_i \) are assumed to be uncorrelated, and have mean 0 and variance \( \sigma^2 \). Then estimates are:

\[
\hat{\beta}_1 = \tilde{b}_1 = \frac{\sum_{i=1}^n X_i Y_i}{\sum_{i=1}^n X_i^2}, \quad \tilde{\text{SSE}} = \sum_{i=1}^n (Y_i - \tilde{b}_1 X_i)^2 = \sum_i Y_i^2 - \tilde{b}_1^2 \sum_i X_i^2.
\]

Also, \( E(\tilde{b}_1) = \beta_1 \), \( E(\tilde{\text{SSE}}) = (n-1)\sigma^2 \), so that \( \tilde{\text{MSE}} = \frac{1}{n-1}\tilde{\text{SSE}} \) is an unbiased estimator of \( \sigma^2 \) and d.f.(\( \tilde{\text{MSE}} \)) = n-1. Var(\( \tilde{b}_1 \)) = \frac{\sigma^2}{\sum_i X_i^2}, \text{ and is estimated by } s^2(\tilde{b}_1) = \frac{\tilde{\text{MSE}}}{\sum_i X_i^2}.

- 100(1 - \( \alpha \))% confidence interval for \( \beta_1 \) : \( (\tilde{b}_1 \pm t(1-\alpha/2; n-1) \cdot s(\tilde{b}_1)) \).
- Estimate of mean response for \( X = X_h \) : \( \tilde{Y}_h = \tilde{b}_1 X_h \) with estimated standard error \( s(\tilde{Y}_h) = \sqrt{\tilde{\text{MSE}} \frac{X_h^2}{\sum_i X_i^2}} \).
- 100(1 - \( \alpha \))% confidence interval for mean response : \( (\tilde{Y}_h \pm t(1-\alpha/2; n-1) \cdot s(\tilde{Y}_h)) \).
Y_h)

• ANOVA decomposition: \(\tilde{SSTO} = \tilde{SSR} + \tilde{SSE}\), where \(\tilde{SSTO} = \sum_i Y_i^2\), with d.f. \(\text{d.f.}(\tilde{SSTO}) = n\), \(\tilde{SSR} = \tilde{b}_1^2 \sum_i X_i^2\) with d.f. \(\text{d.f.}(\tilde{SSR}) = 1\). Reject \(H_0: \beta_1 = 0\) if F-ratio \(F^* = \frac{\tilde{MSR}}{\tilde{MSE}} > F(1-\alpha;1,n-1)\).

Inverse prediction, or calibration problem

In some experimental studies it is important to know the value of \(X\) in order to obtain (on an average) a pre-specified value of \(Y\). The following example illustrates such a situation.

<table>
<thead>
<tr>
<th>X</th>
<th>10</th>
<th>15</th>
<th>15</th>
<th>20</th>
<th>20</th>
<th>20</th>
<th>25</th>
<th>25</th>
<th>28</th>
<th>30</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</td>
<td>160</td>
<td>171</td>
<td>175</td>
<td>182</td>
<td>184</td>
<td>181</td>
<td>188</td>
<td>193</td>
<td>195</td>
<td>200</td>
</tr>
</tbody>
</table>

Here \(Y\) = tensile strength of paper, \(X\) = amount (percentage) of hardwood in the pulp.

Want to find \(X_{\text{h(new)}}\) for given value of \(Y_{\text{h(new)}}\).

Estimate \(\hat{X}_{\text{h(new)}} = \frac{Y_{\text{h(new)}} - b_0}{b_1}\). Estimated standard error of prediction is \(s(\hat{X}_{\text{h(new)}})\) where

\[
s^2(\hat{X}_{\text{h(new)}}) = \frac{\text{MSE}}{b_1^2} \left[1 + \frac{1}{n} + \frac{(\hat{X}_{\text{h(new)}} - \overline{X})^2}{\sum_i (X_i - \overline{X})^2}\right].
\]

Then \(100(1 - \alpha)\)% prediction interval for \(X_{\text{h(new)}}\) is given by \(\hat{X}_{\text{h(new)}} \pm t(1-\alpha/2;n-2) s(\hat{X}_{\text{h(new)}})\).

Fitted model: \(\hat{Y} = 143.8244 + 1.8786X\). SSE = 38.8328, SSTO = 1300.9, SSR = 1262.1, \(R^2 = 0.9701\), \(\sum_i (X_i - \overline{X})^2 = 357.6\), MSE = 4.8541, \(\overline{X} = 20.8\), \(\overline{Y} = 182.9\).

Contributors

• Yingwen Li
• Debasis Paul