ST3232: Design and Analysis of Experiments

Chen Zehua

Department of Statistics & Applied Probability

4:00-6:00 pm, Friday, February 5, 2016
A general introduction

- In many experiments, there are certain continuous variables which might affect the response measurement but cannot be simply controlled by blocking methods.
- The effect of such variables cannot be controlled by designs.
- However, their effects can be adjusted to those of the treatments through a regression model. The adjustment is called regression control.
- Regression control only comes into play at the stage of data analysis. The analysis is also called the analysis of covariance.
- Regression control can be applied with any designs, not only to experiments with one factor.
An important situation

In many experiments, it is the change in the mean value of some variable that is under investigation

- The change in body weight in weight-loss programs,
- The changes in blood pressure in treatments for hypertension,
- The change in tumor size in the study of cancer therapies, etc.

In these examples, the pre-treatment measurement has an effect on the changes, when the changes are compared, the comparison must be adjusted for the effects of the pre-treatment measurements.
Parallel regression lines across levels

Suppose, in the experiment with one factor, the factor has $g$ levels and there is a continuous variable $Z$ associated with each experimental unit. The model without adjusting the effect of $Z$ is

$$y_{ij} = \mu_i + \epsilon_{ij}, \quad i = 1, \ldots, g; \quad j = 1, \ldots, n_i,$$

where $\mu_i$ is the expected effect of treatment $i$.

To adjust for the effect of $Z$, we consider two situations: (a) there is no interaction effect between the treatments and $Z$, (b) the interaction exists.

In situation (a), the following regression model is applicable:

$$y_{ij} = \mu_i + \beta Z_{ij} + \epsilon_{ij}, \quad i = 1, \ldots, g; \quad j = 1, \ldots, n_i. \quad (1)$$

The model is called parallel regression model. The inference on the treatment effects is through that on the parameters of model (1).
Least squares estimates

The LSE of the parameters are given by:

\[
\hat{\beta} = \frac{\sum \sum (y_{ij} - \bar{y}_i)(Z_{ij} - \bar{Z}_i)}{\sum \sum (Z_{ij} - \bar{Z}_i)^2},
\]

\[
\hat{\mu}_i = \bar{y}_i - \hat{\beta} \bar{Z}_i
\]

\[
\hat{\sigma}^2 = s^2 = \frac{\sum \sum (y_{ij} - \hat{\mu}_i - \hat{\beta} Z_{ij})^2}{n. - g - 1},
\]

where \( n. = \sum n_i \).

Let

\[
S_Y^2 = \sum \sum (y_{ij} - \bar{y}_i)^2, \quad S_Z^2 = \sum \sum (Z_{ij} - \bar{Z}_i)^2,
\]

\[
S_{YZ} = \sum \sum (y_{ij} - \bar{y}_i)(Z_{ij} - \bar{Z}_i).
\]
The estimated variances and covariances of \( \hat{\mu}_i \)'s are

\[
\text{Var}(\hat{\mu}_i) = s_{ii} = s^2 \left( \frac{1}{n_i} + \frac{(\bar{Z}_i.)^2}{S_Z^2} \right) ,
\]

\[
\text{Cov}(\hat{\mu}_i, \hat{\mu}_k) = s_{ik} = s^2 \frac{\bar{Z}_i \cdot \bar{Z}_k}{S_Z^2} .
\]

Denote \( S = (s_{ik}) \), \( \hat{\mu} = (\hat{\mu}_1, \cdots, \hat{\mu}_g)' \).

### Inference on overall treatment effect

Let \( C = (c_1, \ldots, c_{g-1}) \) be a \( g \times (g - 1) \) matrix of independent contrast vectors. The hypothesis \( H_0 : \mu_1 = \cdots = \mu_g \) is tested by the following statistic

\[
W = \hat{\mu}' C [C' SC]^{-1} C' \hat{\mu} \sim \chi^2_{g-1} ,
\]

asymptotically. If the assumption of normality holds,

\[
F = W / (g - 1) \sim F_{g-1, n-g-1} .
\]
Multiple comparison

- Let $c_k, k = 1, \ldots, K$, be $K$ contrast vectors of interest.
- The test statistics for these contrasts are constructed as
  \[ L_k = \frac{c' \hat{\mu}}{\sqrt{c' Sc}}, \quad k = 1, \ldots, K. \]
- Each $L_k$ follows a $t$-distribution with df $n. - g - 1$.
- To control the overall error rate, Scheffe’s, Tukey’s, Dunnett’s or Bonferroni’s criterion are used in the usual way depending on the nature of the multiple comparison.
Alternative regression model

- Introducing dummy variables for the treatments as usual:

\[ t_i = \begin{cases} 
1, & \text{if treatment } i, \\
0, & \text{otherwise}, 
\end{cases} \quad i = 2, \ldots, g. \]

- An alternative model is given below:

\[
y = \mu_0 + \sum_{i=2}^{g} \xi_i t_i + \beta Z + \epsilon, \tag{2}
\]

where \( \xi_i = \mu_i - \mu_1 \).
Inference using the alternative regression model

- The null hypothesis $H_0: \mu_1 = \cdots = \mu_g$ is equivalent to
  \[ H_0: \xi_2 = \xi_3 = \cdots = \xi_g = 0. \]

- The contrast $\sum_{i=1}^g c_i \mu_i$ is equivalent to $\sum_{i=2}^g c_i \xi_i$.

- The test statistics are obtained in the form of Wald test statistic in the usual way.

- Multiple comparisons among $\mu_i$’s are made by expressing the corresponding contrasts as linear combinations of the $\xi_i$’s.
Un-parallel regression lines across treatments

If there is an interaction effect between the treatments and the covariate, the regression lines for different treatments are non-parallel. The appropriate model for considering interaction is the following:

\[ y_{ij} = \mu_i + \beta_i Z_{ij} + \epsilon_{ij}. \]  

(3)

This model is equivalent to the following model expressed for an individual observation:

\[ y = \mu_0 + \sum_{i=2}^{g} \xi_i t_i + \beta Z + \sum_{i=2}^{g} \gamma_i t_i Z + \epsilon. \]  

(4)

Note that \( \beta_i = \beta + \gamma_i, i = 2, \ldots, g, \) and \( \beta_1 = \beta. \)
Test for interaction

- Both model (3) and (4) can be used for the data analysis. We focus on model (4) in the following.

- The hypothesis of no interaction is equivalent to

  \[ H_0 : \gamma_2 = \gamma_3 = \cdots = \gamma_g = 0. \]

- Let \( \hat{\gamma} = (\hat{\gamma}_2, \cdots, \hat{\gamma}_g)' \) be the least squares estimates and \( \hat{\Sigma}_\gamma \) be the estimated variance-covariance matrix of \( \hat{\gamma} \). The \( H_0 \) can be tested by Wald statistic:

  \[ W = \hat{\gamma}' \hat{\Sigma}_\gamma^{-1} \hat{\gamma}. \]

- The Wald statistic follows an asymptotic \( \chi^2 \)-distribution with \( \text{df} \) \( g - 1 \). If normality can be assumed for the model, \( F = W / (g - 1) \) follows an exact \( F \)-distribution with \( \text{df} \) \( g - 1 \) and \( n. - 2g \).
Comparison of regression lines

- The $g$ fitted regression lines are given as follows:

$$
\hat{f}_1(Z) = \hat{\mu}_0 + \hat{\beta}Z,
\hat{f}_i(Z) = \hat{\mu}_0 + \hat{\xi}_i + (\hat{\beta} + \hat{\gamma}_i)Z,
$$

$i = 2, \ldots, g$.

- Denote

$$
\hat{f}(Z) = (\hat{f}_1(Z), \hat{f}_2(Z), \ldots, \hat{f}_g(Z))',
\hat{\xi} = (\hat{\xi}_2, \ldots, \hat{\xi}_g)'.
$$

Let $\tilde{c} = (c_1, c')'$ be any contrast vector of dimension $g$. Denote the contrast among the $g$ regression lines by $C(Z)$. We have

$$
C(Z) = \tilde{c}'\hat{f}(Z) = c'(\hat{\xi} + \hat{\gamma}Z).
$$
For a fixed $Z$, the variance of $C(Z)$ is given by

$$\text{Var}(C(Z)) = c'(\hat{\Sigma}_\xi + Z^2\hat{\Sigma}_\gamma + 2Z\hat{\Sigma}_{\xi\gamma})c,$$

where $\hat{\Sigma}_\xi$, $\hat{\Sigma}_\gamma$ and $\hat{\Sigma}_{\xi\gamma}$ are variance matrices of $\hat{\xi}$ and $\hat{\gamma}$ and their covariance matrix respectively.

A confidence band with overall confidence level $1 - \alpha$ is given by

$$C(Z) \pm \sqrt{\text{Var}(C(Z))} \sqrt{2F_{2,n-2g,\alpha}}.$$

(To appreciate the above result, note that

$$C(Z) = (1, Z) \left( \begin{array}{c} c' \hat{\xi} \\ c' \hat{\gamma} \end{array} \right),$$

where $c$ is fixed and $Z$ is arbitrary. An argument similar to that leading to Scheffe’s criterion yields the above result.)
Analysis of variance

By using the R function `lm`, the ANOVA table for unparallel regression model is obtained as

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Covariate</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Treatment by Covariate</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Residuals</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The ANOVA table for parallel regression model is similar but without the entry for interaction. The overall interaction effect and treatment effect can be inferred from the ANOVA table.
Example 8.1

An experiment was conducted to determine if there is a difference in the strength of a monofilament fiber produced by three different machines. The breaking strength (y in bounds) and diameter of the fiber (x in 10^{-3} inches) are given below:

<table>
<thead>
<tr>
<th>Machine 1</th>
<th>Machine 2</th>
<th>Machine 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>y</td>
<td>x</td>
<td>y</td>
</tr>
<tr>
<td>36</td>
<td>20</td>
<td>40</td>
</tr>
<tr>
<td>41</td>
<td>25</td>
<td>48</td>
</tr>
<tr>
<td>39</td>
<td>24</td>
<td>39</td>
</tr>
<tr>
<td>42</td>
<td>25</td>
<td>45</td>
</tr>
<tr>
<td>49</td>
<td>32</td>
<td>44</td>
</tr>
</tbody>
</table>
Example 8.1 (cont.)

- ANOVA table of the interaction model:

<table>
<thead>
<tr>
<th></th>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>machine</td>
<td>2</td>
<td>140.400</td>
<td>70.200</td>
<td>25.0231</td>
<td>0.0002107 ***</td>
</tr>
<tr>
<td>x</td>
<td>1</td>
<td>178.014</td>
<td>178.014</td>
<td>63.4538</td>
<td>2.291e-05 ***</td>
</tr>
<tr>
<td>machine:x</td>
<td>2</td>
<td>2.737</td>
<td>1.369</td>
<td>0.4878</td>
<td>0.6292895</td>
</tr>
<tr>
<td>Residuals</td>
<td>9</td>
<td>25.249</td>
<td>2.805</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- The F-statistic for testing interaction is read from the ANOVA table as $F = 0.4868$ with a $p$-value 0.6293. The interaction is not significant at any reasonable level.

- Note: it can be verified that the F value can be obtained by dividing the Wald statistic with 2.

- A parallel regression model is in order.
Example 8.1 (cont.)

- **ANOVA table** for the model with parallel regression line:

<table>
<thead>
<tr>
<th></th>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>machine</td>
<td>2</td>
<td>140.400</td>
<td>70.200</td>
<td>27.593</td>
<td>5.170e-05 ***</td>
</tr>
<tr>
<td>x</td>
<td>1</td>
<td>178.014</td>
<td>178.014</td>
<td>69.969</td>
<td>4.264e-06 ***</td>
</tr>
<tr>
<td>Residuals</td>
<td>11</td>
<td>27.986</td>
<td>2.544</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- The F-statistic for testing treatment effect read from the ANOVA table is $F = 27.593$ with a $p$-value 0.00005.

- Pairwise comparison statistics:
  - $d_{12} = 1.023592$
  - $d_{13} = -1.430745$
  - $d_{23} = 2.283458$

- Critical value: $c_{0.01} = q_{0.01}(3, 11)/\sqrt{2} = 5.14/\sqrt{2} = 3.6345$.

- Conclusion: none of the pairwise comparisons is significant.