

6. Regression control for prognostic variables

§6.1. Study of measurement of change

Examples of measurement of change:

Change of weight in weight-loss programs;
Change of diastolic or systolic blood pressure due to treatments for hypertension;
Change of tumor size due cancer therapies;
etc..

Example 1: The following table gives summary data of a study comparing two treatments for gingivitis. Each patient was measured before and after treatment on an index of gingivitis, the higher the index, the more severe the gingivitis.

	Treatment	1	2
	n	74	64
Pre- (Z)	Mean	0.6065	0.5578
	sd	0.2541	0.2293
Post- (X)	Mean	0.5514	0.3927
	sd	0.3054	0.1988
Change (D)	Mean	0.0551	0.1651
	sd	0.2192	0.2235

- **A naive approach**

Analysis based on post-treatment measurements:

$$\bar{X}_1 - \bar{X}_2 = 0.5514 - 0.3927 = 0.1587.$$

$$s_x = \sqrt{\frac{73 \times 0.3054^2 + 63 \times 0.1988^2}{73 + 63}}$$

$$= 0.2615.$$

$$t = \frac{0.1587}{0.2615} \sqrt{\frac{74 \times 64}{74 + 64}} = 3.56.$$

The t -statistic has df $n_1 + n_2 - 2 = 136$ and the p-value of the t -test is 0.00013.

Analysis based on the measurement changes:

$$\begin{aligned}\bar{D}_2 - \bar{D}_1 &= (\bar{Z}_2 - \bar{X}_2) - (\bar{Z}_1 - \bar{X}_1) \\ &= (\bar{X}_1 - \bar{X}_2) - (\bar{Z}_1 - \bar{Z}_2) \\ &= 0.1100.\end{aligned}$$

$$\begin{aligned}s_d &= \sqrt{\frac{73 \times 0.2192^2 + 63 \times 0.2235^2}{73 + 63}} \\ &= 0.2212.\end{aligned}$$

$$t = \frac{0.1100}{0.2212} \sqrt{\frac{74 \times 64}{74 + 64}} = 2.91.$$

The t -statistic also has df $n_1 + n_2 - 2 = 136$ and the p-value of the t -test is 0.00106.

Questions:

- Both tests provide evidence for the significance of the treatment difference. Is the evidence reliable?
- Which analysis is more reasonable?
- What factors do we need to consider for such a problem?

Comments:

- It is obvious that the change of measurement should be analyzed instead of the post-treatment measurement, since it is this change that reflects the effect of the treatment. The analysis based on the post-treatment measurement is not valid.
- But the change is not completely due to the treatment effect. For example, extremely higher measurement before treatment tends

to be reduced to a larger extent, i.e., the change is correlated with the pre-treatment measurement. The evidence provided by the analysis based on changes is not very reliable.

- The correlation between the pre-treatment measurement and the measurement change must be taken into account.

• Adjustment for the correlation

Let $\mu_d = E(D)$, $\mu_z = E(Z)$, $\sigma_d^2 = \text{Var}(D)$, $\sigma_z^2 = \text{Var}(Z)$. Let ρ_{zd} be the correlation coefficient between D and Z . Assuming (Z, D) have a joint normal distribution, then given Z , D has normal distribution with mean and variance given below:

$$\begin{aligned}\mu_{d|z} &= \mu_d + \rho_{zd} \frac{\sigma_d}{\sigma_z} (Z - \mu_z), \\ \sigma_{d|z}^2 &= \sigma_d^2 (1 - \rho_{zd}^2).\end{aligned}$$

Then D can be expressed as

$$D = \alpha + \beta_{zd}(Z - \mu_z) + \epsilon.$$

That is, D can be decomposed into three components:

- α : the component caused by treatment;
- $\beta_{zd}(Z - \mu_z)$: the component predictable by Z ;
- ϵ : the component caused by random errors.

Thus, the inference should be made on the adjusted mean differences

$$\bar{D}_1 - \beta_{zd}(\bar{Z}_1 - \mu_z) = \alpha_1, \text{ for treatment 1,}$$

$$\bar{D}_2 - \beta_{zd}(\bar{Z}_2 - \mu_z) = \alpha_2, \text{ for treatment 2.}$$

The raw data of the above example can be described by the model:

$$D_{ij} = \alpha_j + \beta_{zd}(Z_{ij} - \bar{Z}) + \epsilon_{ij},$$

where $j = 1, 2$, for $j = 1, i = 1, \dots, n_1$, for $j = 2, i = 1, \dots, n_2$.

The inference on the significance of difference between two treatments is then equivalent to the inference on the equality of α_1 and α_2 . The problem is reduced to the analysis of a regression model.

Remark:

- Estimate of $\alpha_1 - \alpha_2$ and its estimated variance can be obtained either from the usual regression analysis had the raw data been given or by some ad hoc method using the summary data.
- The test statistic is formed in the usual way as

$$t = \frac{\hat{\alpha}_1 - \hat{\alpha}_2}{\hat{\sigma}_{\hat{\alpha}_1 - \hat{\alpha}_2}}.$$

The test statistic has a t -distribution with

df $n_1 + n_2 - 3$.

- The above t -statistic computed for the example is $t = 3.57$. The p -value of the test is 0.00012.

§6.2. Data analysis with regression control

In many clinical trials, there are certain continuous variables which might affect the end-point measurement but cannot be simply discretized as factors in either completely randomized block designs or stratified randomization designs.

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.

Regression control can be applied with any particular designs. It will be illustrated within the framework of parallel groups designs.

- **Parallel regression lines across treatments**

Model

Suppose there g treatments in a parallel groups design and a continuous variable Z to be controlled. If there is no interaction effect between the treatment and Z , the following regression model is applicable:

$$X_{ij} = \tilde{\mu}_i + \beta(Z_{ij} - \bar{Z}_{..}) + \epsilon_{ij},$$

where $\tilde{\mu}_i$ can be interpreted as the effect of treatment i when Z takes the average value $\bar{Z}_{..}$, $\beta(Z_{ij} - \bar{Z}_{..})$ the component of end-point

measurement predictable by Z regardless of treatments, ϵ_{ij} are i.i.d. errors with mean 0 and variance σ^2 . More conveniently, we can consider the model

$$(1) \quad X_{ij} = \mu_i + \beta Z_{ij} + \epsilon_{ij},$$

with a slightly different interpretation of the parameters. Note that the $\mu_i = \tilde{\mu}_i - \beta \bar{Z}_{i.}$

Estimation

$$\hat{\beta} = \frac{\sum \sum (X_{ij} - \bar{X}_{i.})(Z_{ij} - \bar{Z}_{i.})}{\sum \sum (Z_{ij} - \bar{Z}_{i.})^2},$$

$$\hat{\mu}_i = \bar{X}_{i.} - \hat{\beta} \bar{Z}_{i.},$$

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

where $n. = \sum n_i$.

Let

$$S_X = \sum \sum (X_{ij} - \bar{X}_{i.})^2,$$

$$S_{XZ} = \sum \sum (X_{ij} - \bar{X}_{i.})(Z_{ij} - \bar{Z}_{i.}).$$

The notation S_Z^2 is used to denote the same sum of squares for Z .

Let s_{ik} denote the estimated covariance between $\hat{\mu}_i$ and $\hat{\mu}_k$. The estimated variance of $\hat{\beta}$ and s_{ik} 's are given below:

$$\text{Var}(\hat{\beta}) = s^2 / S_Z^2,$$

$$s_{ii} = s^2 \left(\frac{1}{n_i} + \frac{(\bar{Z}_{i.})^2}{S_Z^2} \right),$$

$$s_{ik} = s^2 \bar{Z}_{i.} \bar{Z}_{k.} / S_Z^2.$$

Let

$$S = (s_{ik}), \quad \hat{\boldsymbol{\mu}} = (\hat{\mu}_1, \dots, \hat{\mu}_g)'$$

Testing the significance of treatment effect

The hypothesis of no treatment effects is expressed as

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_g.$$

The hypothesis can be tested by Wald test. Let C be the matrix of $g - 1$ linearly independent contrast vectors, e.g.,

$$C = \begin{pmatrix} 1 & -1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & -1 & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & 1 & -1 \end{pmatrix}.$$

The Wald statistic is formed as

$$\chi^2 = \hat{\boldsymbol{\mu}}' C' [C S C']^{-1} C \hat{\boldsymbol{\mu}}.$$

Asymptotically, χ^2 has a χ^2 -distribution with df $g - 1$. If normality is assumed for ϵ_{ij} ,

$F = \chi^2 / (g - 1)$ has an exact F -distribution with df $g - 1$ and $n. - g - 1$.

If $\chi^2 > \chi_{g-1, \alpha}^2$ or $F > F_{g-1, n.-g-1, \alpha}$, H_0 is rejected.

Multiple comparison

Let $\mathbf{c}_k, k = 1, \dots, K$, be K contrast vectors of interest. The test statistics for these contrasts are constructed as

$$L_k = \frac{\mathbf{c}'_k \hat{\boldsymbol{\mu}}}{\sqrt{\mathbf{c}'_k S \mathbf{c}_k}}, \quad k = 1, \dots, K.$$

Each L_k following 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.

Example 2: The summary data in the following table are from a study comparing three methods of treating the learning disabilities of children with respiratory diseases. X is the number of correct answers to a test with 15 questions and Z is the number of correct answers to a similar test with 7 questions.

Quantity	Treatment 1	Treatment 2	Treatment 3
n_i	19	20	20
$\bar{X}_i.$	5.3158	8.3000	8.5500
$\sum(X_{ij} - \bar{X}_i.)^2$	67.1053	76.2000	139.9500
$\bar{Z}_i.$	2.3158	2.4500	3.1500
$\sum(Z_{ij} - \bar{Z}_i.)^2$	23.1050	15.9500	22.5500
$\sum(X_{ij} - \bar{X}_i.)(Z_{ij} - \bar{Z}_i.)$	27.1053	18.3000	34.3500

R-code for the computation:

```
s.x = c(67.1053,76.2000,139.9500)
s.z =c(23.1053,15.9500,22.5500)
s.xz =c(27.1053,18.3,34.35)
x.m = c(5.3158,8.3,8.55)
z.m =c(2.3158,2.45,3.15)
n.i =c(19,20,20)
n=sum(n.i)
```

```

ss.x=sum(s.x)
ss.z=sum(s.z)
ss.xz =sum(s.xz)
b = ss.xz/ss.z
mu = x.m -b*z.m
s2=(ss.x-ss.xz^2/ss.z)/(n-3-1)
S = s2*(z.m%*%t(z.m)/ss.z + diag(1/n.i) )
C = matrix(c(-1,1,0,-1,0,1), byrow=T, ncol=3)
V = C%*%S%*%t(C)
F = t(C.u)%*%solve(V)%*%C.u/2

```

It is computed that

$$\hat{\beta} = 1.2946,$$

$$\hat{\boldsymbol{\mu}} = (2.317725, 5.128187, 4.471955),$$

$$s^2 = 3.2728,$$

$$S = \begin{pmatrix} 0.457157 & 0.301415 & 0.387534 \\ 0.301415 & 0.482521 & 0.409992 \\ 0.387534 & 0.409992 & 0.690771 \end{pmatrix},$$

$$F = 12.52398.$$

Point-of-view of analysis of covariance

The F statistic for testing the significance of treatment effects can be obtained from an anova table with the original SS's of the variable X adjusted by Z . The adjusted SS's have the same relationship among themselves as the unadjusted SS's, i.e.,

$$\text{TSS}' = \text{BSS}' + \text{WSS}'.$$

The F statistic is then form as

$$F = \frac{\text{BSS}'(g - 1)}{\text{WSS}'/(n. - g - 1)}.$$

The adjusted SS's are given as follows:

$$\begin{aligned}\text{TSS}' &= \sum \sum (X_{ij} - \bar{X}_{..})^2 - \frac{[\sum \sum (X_{ij} - \bar{X}_{..})(Z_{ij} - \bar{Z}_{..})]^2}{\sum \sum (Z_{ij} - \bar{Z}_{..})^2}, \\ \text{WSS}' &= \sum \sum (X_{ij} - \bar{X}_{i.})^2 - \frac{[\sum \sum (X_{ij} - \bar{X}_{i.})(Z_{ij} - \bar{Z}_{i.})]^2}{\sum \sum (Z_{ij} - \bar{Z}_{i.})^2}, \\ \text{BSS}' &= \text{TSS}' - \text{WSS}'.\end{aligned}$$

An alternative model

Introducing dummy variables for the treatments as usual:

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

The data can then be described for each individual as

$$(2) \quad X = \mu_0 + \sum_{i=2}^g \xi_i t_i + \beta Z + \epsilon,$$

where $\xi_i = \mu_i - \mu_1$.

The null hypothesis of no treatment effect becomes:

$$H_0 : \xi_2 = \xi_3 = \dots = \xi_g = 0.$$

The hypothesis is tested by Wald test or F -test in the usual way.

Multiple comparisons are made by expressing contrasts as linear combinations of the ξ_j 's.

- **Un-parallel regression lines across treatments**

The model

If there is an interaction effect between the treatments and the variable to be controlled, the regression lines for different treatments will be non-parallel. The appropriate model is then

$$(3) \quad X_{ij} = \mu_i + \beta_i Z_{ij} + \epsilon_{ij}.$$

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

$$(4) \quad X = \mu_0 + \sum_{i=2}^g \xi_i t_i + \beta Z + \sum_{i=2}^g \gamma_i t_i Z + \epsilon.$$

Note that $\beta_i = \beta + \gamma_i, i = 2, \dots, g$, 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 $\hat{\boldsymbol{\gamma}} = (\hat{\gamma}_2, \cdots, \hat{\gamma}_g)'$ be the least squares estimates and $\hat{\Sigma}_{\boldsymbol{\gamma}}$ be the estimated variance-covariance matrix of $\hat{\boldsymbol{\gamma}}$. (This matrix can be directly extract from the variance-covariance matrix of the fitted object obtained by the R-function `lm`).

The H_0 can be tested by Wald statistic:

$$\chi^2 = \hat{\boldsymbol{\gamma}}' \hat{\Sigma}_{\boldsymbol{\gamma}}^{-1} \hat{\boldsymbol{\gamma}}.$$

The statistic follows an asymptotic χ^2 -distribution with df $g - 1$. If normality can be assumed for

the model, $F = \chi^2/(g - 1)$ follows an exact F -distribution with df $g - 1$ and $n. - 2g$.

Comparison of regression lines

The g fitted regression lines are given as follows:

$$\begin{aligned}\hat{f}_1(Z) &= \hat{\mu}_0 + \hat{\beta}Z, \\ \hat{f}_i(Z) &= \hat{\mu}_0 + \hat{\xi}_i + (\hat{\beta} + \hat{\gamma}_i)Z, \\ i &= 2, \dots, g.\end{aligned}$$

Denote

$$\begin{aligned}\hat{\mathbf{f}}(Z) &= (\hat{f}_1(Z), \hat{f}_2(Z), \dots, \hat{f}_g(Z))', \\ \hat{\boldsymbol{\xi}} &= (\hat{\xi}_2, \dots, \hat{\xi}_g)'. \end{aligned}$$

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

$$C(Z) = \tilde{\mathbf{c}}' \hat{\mathbf{f}}(Z) = \mathbf{c}'(\hat{\boldsymbol{\xi}} + \hat{\boldsymbol{\gamma}}Z).$$

For a fixed Z , the variance of $C(Z)$ is given by

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

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

$$C(Z) = (1, Z) \begin{pmatrix} \mathbf{c}'\hat{\xi} \\ \mathbf{c}'\hat{\gamma} \end{pmatrix},$$

where \mathbf{c} is fixed and Z is arbitrary. An argument similar to that leading to Scheffe's criterion yields the above result.)

§6.3. More complicated designs with regression control

- The regression control can be applied with any designs.
- More than one covariates can be adjusted in the same way as in the case of one covariate.
- **The general method:** For whatever design, the adjustment is done by imposing the covariate terms on top of the model for that design.

Example 3 : The following table provides additional information on a covariate Z for the first 10 pairs of the patient in a randomized block design considered in Lecture notes 4.

Pair	Imipramine		Placebo	
	<i>Z</i>	<i>X</i>	<i>Z</i>	<i>X</i>
1	27	6	18	4
2	18	4	21	7
3	25	6	24	12
4	23	7	24	10
5	22	5	18	2
6	24	6	27	11
7	25	8	21	9
8	22	7	21	5
9	26	8	26	11
10	19	3	20	8

The following R-code is used for the computation:

```
x=c(6,4,6,7,5,6,8,7,8,3,4,7,12,10,2,11,9,5,11,8)
z=c(27,18,25,23,22,24,25,22,26,19,18,21,24,24,18,27,21,21,26,20)
pair=factor(rep(c(1:10),2))
tmt =factor(c(rep(1,10), rep(2,10)))
cbind(x,z,pair,tmt)
options(contrasts=c("contr.treatment","contr.poly"))
lm.fit=lm(x~z+pair+tmt)
anova(lm.fit)
```

The computation yields the anova table:

Analysis of Variance Table

Response: x

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
z	1	62.724	62.724	20.3760	0.001965	**
pair	9	25.526	2.836	0.9214	0.551509	
tmt	1	30.073	30.073	9.7695	0.014108	*
Residuals	8	24.627	3.078			

To test the adjusted treatment effects, the F -ratio is formed by the mean sum of square of the treatments and the mean residual sum of squares as

$$F = \frac{30.073}{3.078} = 9.7695.$$

With the distribution $F_{1,8}$, the p -value is 0.0141.

Note: the df is obtained as the total number of observations (20) minus the number of parameters in the model (1 for interception, 1 for the covariate, 9 for pairs and 1 for treatment).