Chapter 25

Analysis of Covariance

Analysis of covariance (or ANCOVA) introduces regression elements into an ANOVA analysis. This is done primarily to increase the precision of the ANOVA analysis.

25.1 Basic Ideas

SAS program: att7-25-1-drugs-ANCOVA,plots

Exercise 25.1 (Basic Ideas)

1. *Analysis of variance (review): patient response versus drug types.*
   Fifteen patients responses are recorded for three drug types.

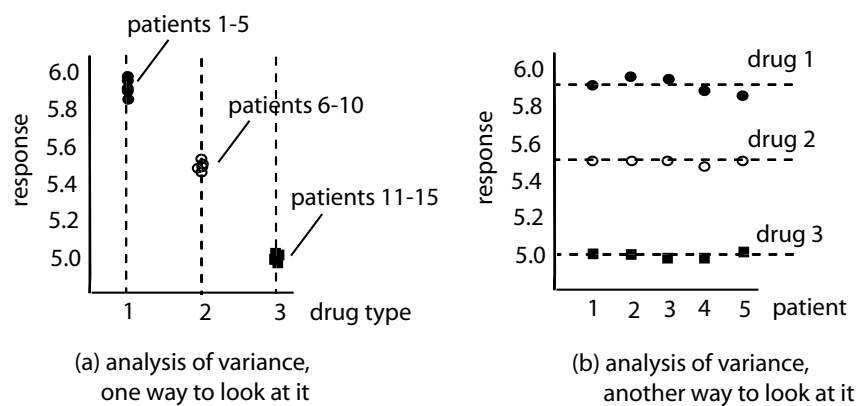| drug 1	5.90	5.92	5.91	5.89	5.88	$\overline{Y}_{1.} \approx 5.90$
drug 2	5.50	5.50	5.50	5.49	5.50	$\overline{Y}_{2.} = 5.50$
drug 3	5.01	5.00	4.99	4.98	5.02	$\overline{Y}_{3.} \approx 5.00$

Figure 25.1 (ANOVA: Patient response versus drug types)

(a) *Figure (a) analysis of variance*

    i. The point $(x, y) = (1, 5.90)$ corresponds to the first patient response of 5.90 to drug 1, where $x = 1$ refers to
(choose one) **patient** / **drug type**

    ii. The x–axis, "drug type", is a
(choose one) **qualitative** / **quantitative** variable.

    iii. **True** / **False**
It is *not* possible to plot a drug response at an x–value of $x = 1.5$, say, because there is no "drug type 1.5".

    iv. There are (choose one) **five** / **fifteen** patients in this study.

    v. **True** / **False**
In an analysis of variance (ANOVA) of this data, we are interested primarily in whether the average patient responses to the three different drugs is the same or different.

    vi. There are *two* (and only two) variables in this study, including
(*choose none, one or more!*)
      A. patient response to drug
      B. drug type (1, 2 and 3)
      C. length of previous illness

(b) *Figure (b) analysis of variance*

    i. The x–axis, "patient", is a
(choose one) **qualitative** / **quantitative** variable.

    ii. The point $(x, y) = (1, 5.90)$ corresponds to the first patient response of 5.90 to drug 1, where $x = 1$ refers to
(choose one) **patient** / **drug type**.

    iii. **True** / **False**
It is *not* possible to plot a drug response at an x–value of $x = 1.5$, say, because there is no "patient 1.5".

    iv. There are (choose one) **five** / **fifteen** patients. Figures (a) and (b) give the same information in slightly different ways.

    v. The three horizontal dotted lines correspond to (choose one)
      A. the three average patient responses,
$\overline{Y}_{1.} \approx 5.90$, $\overline{Y}_{2.} \approx 5.50$, $\overline{Y}_{3.} \approx 5.00$.
      B. the three linear regressions of the three drug responses regressed on patient. (Is it possible to regress on a qualitative variable?)

    vi. There are *two* (and only two) variables in this study, including
(*choose none, one or more!*)
      A. patient response to drug
      B. drug type (1, 2 and 3)

C. length of previous illness

2. *Analysis of covariance: patient response versus drug types, with length of previous illness as a concomitant variable.*
   Patients responses are recorded for three drug types, as well as the concomitant variable: length of time (in hours) they were ill before receiving any of the three drugs.

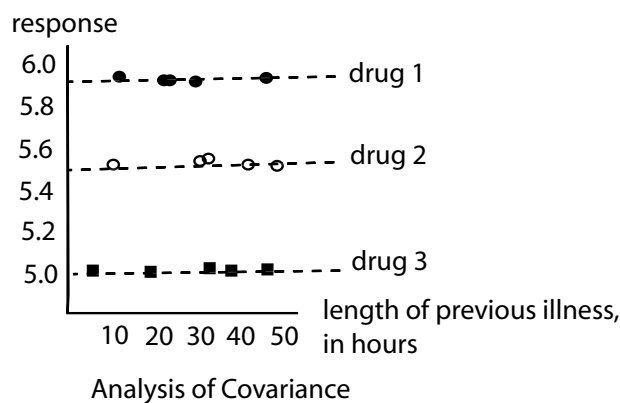| drug 1	5.90	5.92	5.91	5.89	5.88
time previously ill	23	10	45	22	29
drug 2	5.50	5.50	5.50	5.49	5.50
time previously ill	33	31	10	51	43
drug 3	5.01	5.00	4.99	4.98	5.02
time previously ill	05	47	19	38	33

Figure 25.2 (ANCOVA: Patient response versus drug types, with concomitant, length of previous illness))

(a) **True / False** The point $(x, y) = (23, 5.90)$ corresponds to a response of 5.90 for a patient who had been ill for 23 hours before receiving drug 1.

(b) It *is* possible to plot a drug response at an x–value of $x = 13.3$, say, because it is possible for a patient to have been 13.3 hours ill before receiving a drug. The x–axis of the plot, "length of previous illness", is (choose one) **quantitative / qualitative**. The variable, "length of previous illness" is called a *concomitant* ("regression") variable.

(c) The three (more or less) horizontal dotted lines correspond to (choose one)
   i. the three averages, $\overline{Y}_{1.} \approx 5.90$, $\overline{Y}_{2.} \approx 5.50$, $\overline{Y}_{3.} \approx 5.00$.
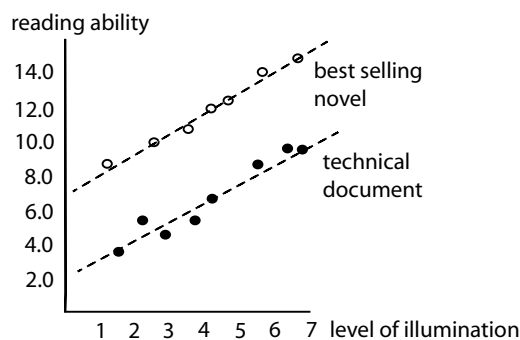   ii. the three linear regressions of the three drug responses regressed on "length of previous illness".

(d) There are *three* variables in this study, including
(*choose none, one or more!*)

   i. patient response to drug

  ii. drug type (1, 2 and 3)

 iii. length of previous illness

The patient response is not only related to the *drug type* in an "ANOVA" kind of way, but also, the patient response is related to the *length of previous illness* in a "linear regression" kind of way.

3. *Another analysis of covariance: reading ability versus type of reading material, with level of illumination as a concomitant variable.*
Reading ability is recorded for two types of reading material: best–selling novel and technical document. In addition, the *concomitant* variable, level of illumination, is also recorded.

technical document	3.6	5.3	4.3	5.2	6.6	8.5	9.6	9.5
level of illumination	1.5	2.2	2.9	3.8	4.1	5.5	6.4	6.8
best–selling novel	8.6	9.8	10.6	11.9	12.2	13.9	14.8	
level of illumination	1.2	2.5	3.5	4.1	4.6	5.6	6.6	

Figure 25.3 (ANCOVA: Reading ability versus type of reading material, with concomitant, level of illumination))

(a) **True / False** The point $(x, y) = (1.5, 3.6)$ corresponds to a reading ability of 3.6 at a level of illumination of 1.5 for a subject who reads a technical document.

(b) The x–axis of the plot, "level of illumination", is
(choose one) **quantitative / qualitative**.

    (c) The concomitant variable is
       (choose one) **level of illumination / type of reading material**.

    (d) The two dotted lines correspond to (choose one)

        i. the two averages, $\overline{Y}_{1\cdot}$, and $\overline{Y}_{2\cdot\cdot}$.

        ii. the two regressions (novel and technical document) of the reading ability regressed on "level of illumination".

    (e) There are three variables in this study, including
       (*choose none, one or more!*)

        i. reading ability

        ii. reading material type (1: novel, 2: technical document)

        iii. level of illumination

    (f) A "level of illumination" concomitant variable is one which
       (*choose none, one or more!*)

        i. influences the "reading ability" variable.

        ii. is *not* influenced by and does not influence the "type of reading material" variable (factor, treatment).

        iii. the range of the concomitant variable is more or less the same for both the novel (1.5 to 6.8) and technical document levels (1.2 to 6.6).

    (g) In general, a "good" concomitant variable is one which
       (*choose none, one or more!*)

        i. influences the dependent response variable.

        ii. is *not* influenced by and does not influence the other qualitative variable (factor, treatment).

        iii. the range of the concomitant variable is the same for all levels of the treatment.

25.2   Single–Factor Covariance Model

<div align="center">SAS program: att7-25-2-drugs-ANCOVA</div>

Exercise 25.2 (Single–Factor Covariance Model)
We look at the model for a single–factor covariance (ANCOVA) and some of the related properties.

1. *General analysis of covariance: reading ability versus type of reading material, with level of illumination as a concomitant variable.*
   Reading ability is recorded for two types of reading material: best–selling novel and technical document. In addition, the *concomitant* variable, level of illumination, is also recorded.

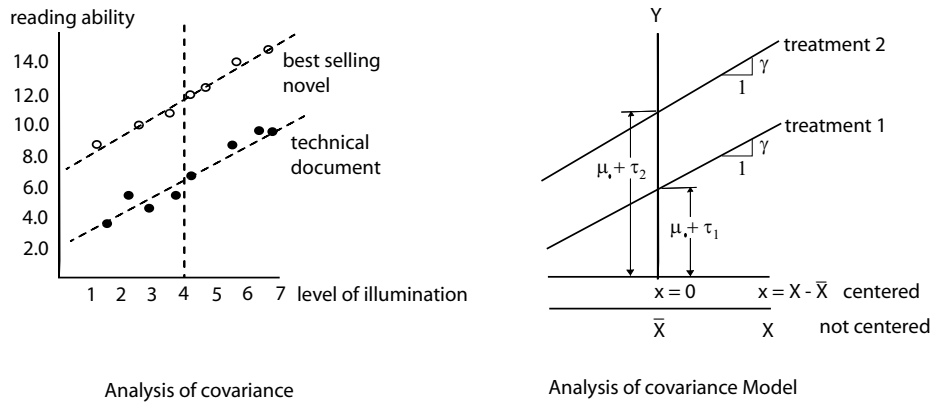| technical document	3.6	5.3	4.3	5.2	6.6	8.5	9.6	9.5
level of illumination	1.5	2.2	2.9	3.8	4.1	5.5	6.4	6.8
best–selling novel	8.6	9.8	10.6	11.9	12.2	13.9	14.8	
level of illumination	1.2	2.5	3.5	4.1	4.6	5.6	6.6	

Figure 25.4 (ANCOVA: Example and model))

(a) *One possible ANCOVA model.*
One *possible* factor effects model for ANCOVA could be given by

$$Y_{ij} = \mu. + \tau_i + \gamma X_{ij} + \varepsilon_{ij}$$

where Y_{ij} is the dependent (response) variable (choose one)

  i. reading ability

  ii. reading material type (1: novel, 2: technical document)

 iii. level of illumination

  iv. slope: change in reading ability for a unit change in illumination

and τ_i is the qualitative treatment parameter (choose one)

  i. reading ability

  ii. reading material type (1: novel, 2: technical document)

 iii. level of illumination

  iv. slope: change in reading ability for a unit change in illumination

and X_{ij} is the concomitant quantitative variable (choose one)

  i. reading ability

  ii. reading material type (1: novel, 2: technical document)

 iii. level of illumination

  iv. slope: change in reading ability for a unit change in illumination

and γ is the slope parameter measuring the (choose one)

  i. reading ability

  ii. reading material type (1: novel, 2: technical document)

  iii. level of illumination

  iv. change in reading ability for a unit change in illumination

(b) *Estimating the ANCOVA model*

From SAS, a possible factor effects model for ANCOVA,

$$Y_{ij} = \mu. + \tau_i + \gamma X_{ij} + \varepsilon_{ij}$$

is estimated by (choose one)

  i. $\hat{Y} = 4.311 - 2.635 I_1 + 1.180 X$

  ii. $\hat{Y} = 9.140 - 2.635 I_1 + 1.180 x$

where

$$I_1 = \begin{cases} 1, & \text{if case from technical document} \\ -1, & \text{if case from best–selling novel,} \end{cases}$$

(c) *Using the ANCOVA model*

From SAS, the possible factor effects model for ANCOVA,

$$Y_{ij} = \mu. + \tau_i + \gamma X_{ij} + \varepsilon_{ij}$$

is estimated by $\hat{Y} = 4.311 - 2.635 I_1 + 1.180 X$.

Consequently, a subject who reads a best–selling novel ($I_1 = -1$, not 1!) with level of illumination $X = 3.5$ would be expected (estimated) to have a reading ability of
$\hat{Y} = 4.311 - 2.635(-1) + 1.180(3.5) =$ (choose one) **4.932 / 5.806 / 11.076**

A subject who reads a technical novel ($I_1 = 1$) with level of illumination $X = 3.5$ would be expected to have a reading ability of
$\hat{Y} = 4.311 - 2.635(1) + 1.180(3.5) =$ (choose one) **4.932 / 5.806 / 6.234**

A subject who reads a technical novel ($I_1 = -1$) with the *overall average* level of illumination, $X = \bar{X}.. = 4.0833$, would be expected to have a reading ability of
$\hat{Y} = 4.311 - 2.635(-1) + 1.180(4.0833) =$
(choose one) **4.932 / 6.494 / 11.764**

A subject who reads a best–selling novel ($I_1 = 1$) with $X = \bar{X}.. = 4.0833$, would be expected to have a reading ability of
$\hat{Y} = 4.311 - 2.635(1) + 1.180(4.0833) =$
(choose one) **4.494 / 5.806 / 11.764**

(d) *Another (centered) ANCOVA model*
    Another (centered) factor effects model for ANCOVA is given by

    $$Y_{ij} = \mu_. + \tau_i + \gamma(X_{ij} - \bar{X}_{..}) + \varepsilon_{ij}$$
    $$= \mu_. + \tau_i + \gamma x_{ij} + \varepsilon_{ij}$$

    where

    $$x_{ij} = X_{ij} - \bar{X}_{..}$$

    is estimated by (choose one)

    i. $\hat{Y} = 4.311 - 2.635I_1 + 1.180X$
    ii. $\hat{Y} = 9.140 - 2.635I_1 + 1.180x$

    where

    $$I_1 = \begin{cases} 1, & \text{if case from technical document} \\ -1, & \text{if case from best–selling novel,} \end{cases}$$

(e) *Using the centered ANCOVA model*
    From SAS, the centered factor effects model for ANCOVA,

    $$Y_{ij} = \mu_. + \tau_i + \gamma x_{ij} + \varepsilon_{ij}$$

    is estimated by $\hat{Y} = 9.140 - 2.635I_1 + 1.180x$.

    A subject who reads a technical novel ($I_1 = 1$)
    with the *overall average* level of illumination,

    $$x = X - \bar{X}_{..} = \bar{X}_{..} - \bar{X}_{..} = 4.0833 - 4.0833 = 0$$

    would be expected to have a reading ability of
    $\hat{Y} = 9.140 - 2.635(1) + 1.180(0) =$
    (choose one) **4.932 / 6.505 / 11.323**
    which, notice, is the same (to within round–off error) answer that was
    obtained above using the *un*–centered ANCOVA above.

    A subject who reads a best–selling novel ($I_1 = 1$) where $x = 0$
    would be expected to have a reading ability of
    $\hat{Y} = 9.140 - 2.635(1) + 1.180(0) =$
    (choose one) **4.494 / 5.806 / 11.775**
    which, notice, is the same answer as was obtained before.

(f) *Comparing un–centered ANCOVA with centered ANCOVA model.*
    The centered ANCOVA,

    $$Y_{ij} = \mu_. + \tau_i + \gamma x_{ij} + \varepsilon_{ij}$$

    where $x = X_{ij} - \bar{X}_{..}$ is a set of linear regressions where the various mean
    treatments, $\mu_. + \tau_i$ (choose one) **appear / do not appear** as the y–
    intercepts of these regressions.
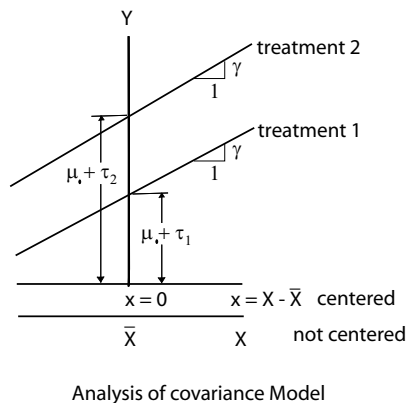
2. *Inference For Treatment Effects.*

Analysis of covariance Model

Figure 25.5 (Analysis of Covariance Model))

True / False In the factor effects model for ANCOVA,

$$Y_{ij} = \mu. + \tau_i + \gamma(X_{ij} - \bar{X}..) + \varepsilon_{ij},$$

the y–intercept (when $X_{ij} - \bar{X}.. = 0$) is $\mu. + \tau_i$. In other words, the difference between two treatment means is, for example,

$$\mu. + \tau_1 - (\mu. + \tau_1) = \tau_1 - \tau_2$$

and a possible related test might be

$$H_0: \ \tau_1 = \tau_2 = 0 \ \ vs \ \ H_a: \ \tau_i \neq 0, \ \ i = 1, 2$$

3. *Slopes Must Be Parallel.*
   True / False The regression lines for all the treatment means must be parallel to, for example, test the difference between two treatment means,

$$\mu. + \tau_1 - (\mu. + \tau_1) = \tau_1 - \tau_2,$$

The observed data must be checked to see if it fits the parallel–slope requirement before being able to use the covariance model.

4. *Regression Formulation of Covariance Model.*
   The regression formulation of the covariance model is given by

$$Y_{ij} = \mu. + \tau_1 I_{ij1} + \cdots + \tau_{r-1} I_{ij,r-1} + \gamma(X_{ij} - \bar{X}..) + \varepsilon_{ij},$$

   where

$$I_1 = \begin{cases} 1, & \text{if case from treatment 1} \\ -1, & \text{if case from treatment r} \\ 0, & \text{otherwise,} \end{cases}$$

$\vdots$

$$I_{r-1} = \begin{cases} 1, & \text{if case from treatment r - 1} \\ -1, & \text{if case from treatment r} \\ 0, & \text{otherwise.} \end{cases}$$

where there are (choose one) $\boldsymbol{r-1}$ / $\boldsymbol{r}$ / $\boldsymbol{r+1}$ indicator variables, I_i, for the r levels of the treatment variable, τ_i, $i = 1, \ldots, r$.

5. *Generalization of Covariance Model.*
   Identify which of the following are possible generalizations to the covariance model.

   (a) The concomitant X could be random as opposed to fixed, as it is now[1].

   (b) There could be a nonlinear relationship, such as

   $$Y_{ij} = \mu_{\cdot} + \tau_i + \gamma(X_{ij} - \bar{X}_{\cdot\cdot})^2 + \varepsilon_{ij},$$

   (c) There might be more than one concomitant variable, such as

   $$Y_{ij} = \mu_{\cdot} + \tau_i + \gamma_1(X_{ij1} - \bar{X}_{\cdot\cdot 1}) + \gamma_2(X_{ij2} - \bar{X}_{\cdot\cdot 2}) + \varepsilon_{ij}$$

   (d) The treatment and concomitant variable might interact with one another,

   $$Y_{ij} = \mu_{\cdot} + \tau_i + \tau_i\gamma(X_{ij1} - \bar{X}_{\cdot\cdot}) + \varepsilon_{ij}$$

6. *Checking the Assumptions For the Covariance Model.*
   Identify which of the following need to be checked to allow the use of the covariance model.

   (a) normality of error terms

   (b) equality of error variances for different treatments

   (c) linearity of regression relation with concomitant variables

   (d) uncorrelatedness of error terms

   Also, although not a requirement to allow the use of the covariance model, to allow inference for the treatment parameters in a covariance model, the slopes of the different treatment regression lines need to be shown to be parallel to one another.

---

[1]This is analogous to allowing random, as opposed to fixed, effects in the ANOVA model.

25.3 Example of Single–Factor Covariance Analysis

SAS program: att7-25-3-drug,read-ANCOVA,inference

Exercise 25.3 (Example of Single–Factor Covariance Model)

1. *Analysis of Covariance: Reading Ability Versus Type of Reading Material, With Level of Illumination As A Concomitant Variable.*
   Reading ability is recorded for two types of reading material: best–selling novel and technical document. In addition, the *concomitant* variable, level of illumination, is also recorded.

technical document	3.6	5.3	4.3	5.2	6.6	8.5	9.6	9.5
level of illumination	1.5	2.2	2.9	3.8	4.1	5.5	6.4	6.8
best–selling novel	8.6	9.8	10.6	11.9	12.2	13.9	14.8	
level of illumination	1.2	2.5	3.5	4.1	4.6	5.6	6.6	

   (a) *Check the ANCOVA assumptions.*

      i. *Residual plot*
         From SAS, the residual plot indicates (choose one)

         A. the variance for the two difference treatments (technical document and best–selling novel) are the *same*.

         B. the variance for the two difference treatments (technical document and best–selling novel) are *different*.

      ii. *Normal probability plot*
         The normal probability plot (choose one)

         A. is linear and so indicates the residuals are normal.

         B. is linear, except for one outlier, and so indicates residuals are *not* normal.

      Even though the residual plots indicate that the data does *not* satisfy the ANCOVA assumptions, we will continue with the analysis in any case.

   (b) *Regression Formulation, For When Lines Parallel.*
      The regression formulation of the *full* (interaction) covariance model is given by
      $$Y_{ij} = \mu_. + \tau_1 I_{ij1} + \gamma x_{ij} + \beta_1 I_{ij1} x_{ij} + \varepsilon_{ij}$$

      and, from SAS, estimated by (choose one)

         i. $\hat{Y} = 9.1407 - 2.6357 I_1 + 1.1825 x - 0.0148 I_1 x$
         ii. $\hat{Y} = 9.1396 - 2.6354 I_1 + 1.1805 x$