The coverage probability of confidence intervals in one-way analysis of covariance after two F tests

Waruni Abeysekera

Department of Mathematics and Statistics
La Trobe University, Melbourne, Australia

December 02, 2013

Introduction

Volume 3 of *Analysis of Messy Data* by Milliken & Johnson (2002) provides detailed recommendations about sequential model development for the analysis of covariance.

Introduction

Volume 3 of *Analysis of Messy Data* by Milliken & Johnson (2002) provides detailed recommendations about sequential model development for the analysis of covariance.

In his review of this volume, Koehler (2002) *JASA* asks whether users should be concerned about the effect of this sequential model development on the coverage probabilities of confidence intervals for comparing treatments.

Introduction

Volume 3 of *Analysis of Messy Data* by Milliken & Johnson (2002) provides detailed recommendations about sequential model development for the analysis of covariance.

In his review of this volume, Koehler (2002) *JASA* asks whether users should be concerned about the effect of this sequential model development on the coverage probabilities of confidence intervals for comparing treatments.

We present a general methodology for the examination of these coverage probabilities in the context of the two-stage model selection procedure that uses two F tests and is proposed in Chapter 2 of this volume.

We present a general methodology for the examination of these coverage probabilities in the context of the two-stage model selection procedure that uses two F tests and is proposed in Chapter 2 of this volume.

Our conclusion is that users should be very concerned about the coverage probabilities of confidence intervals for comparing treatments constructed after this two-stage model selection procedure.

The one-way analysis of covariance model described in Chapter 2 of Milliken & Johnson (2002)

Consider the following one-way analysis of covariance model described in Chapter 2 of Milliken & Johnson (2002).

$$Y_{ij} = a_i + b_i \left(x_{ij} - \bar{x}\right) + \varepsilon_{ij} \tag{1}$$

where Y_{ij} is the response of the j'th experimental unit ($j = 1, ..., n_i$) receiving treatment i ($i = 1, ..., k$), when the covariate takes the value x_{ij}.

The ε_{ij} are independent and identically $N(0, \sigma^2)$ distributed, where σ^2 is an unknown positive parameter.

The a_i and the slopes b_i are unknown parameters.

The two-stage model selection procedure proposed in Chapter 2 of Milliken & Johnson (2002)

Milliken & Johnson (2002, Section 2.3) propose the following two-stage procedure to determine the form of the model.

The two-stage model selection procedure proposed in Chapter 2 of Milliken & Johnson (2002)

Milliken & Johnson (2002, Section 2.3) propose the following two-stage procedure to determine the form of the model.

Stage 1: test the null hypothesis that the slopes b_i are all zero against the alternative hypothesis that they are not all zero.

If this null hypothesis is accepted then assume that the slopes b_i are all zero; otherwise proceed to Stage 2.

The two-stage model selection procedure proposed in Chapter 2 of Milliken & Johnson (2002)

Milliken & Johnson (2002, Section 2.3) propose the following two-stage procedure to determine the form of the model.

Stage 1: test the null hypothesis that the slopes b_i are all zero against the alternative hypothesis that they are not all zero.

If this null hypothesis is accepted then assume that the slopes b_i are all zero; otherwise proceed to Stage 2.

Stage 2: test the null hypothesis that the slopes b_i are all equal against the alternative hypothesis that they are not all equal.

If this null hypothesis is accepted then assume that the slopes b_i are all equal; otherwise this assumption is not made.

Confidence intervals constructed after the two-stage model selection procedure

Suppose that the parameter of interest θ is a specified linear contrast of the expected responses, for a given value of the covariate, which can be expressed as,

$$\theta = \mathbf{a}^\top \boldsymbol{\beta}$$

where $\boldsymbol{\beta} = (a_1, \ldots, a_k, b_1, \ldots, b_k)$ and $\mathbf{a}$ is the vector of contrast coefficients.

Confidence intervals constructed after the two-stage model selection procedure

Suppose that the parameter of interest θ is a specified linear contrast of the expected responses, for a given value of the covariate, which can be expressed as,

$$\theta = \mathbf{a}^\top \boldsymbol{\beta}$$

where $\boldsymbol{\beta} = (a_1, \ldots, a_k, b_1, \ldots, b_k)$ and $\mathbf{a}$ is the vector of contrast coefficients.

The CI for θ has three different forms, depending on the model resulting from the two-stage procedure.

1. When the slopes b_i are assumed to be all zero.
2. When the slopes b_i are assumed to be all equal.
3. When fitting the full model.

To find the CP of this CI, we use the law of total probability.

To find the CP of this CI, we use the law of total probability.

We derive a simplified expression for this CP and show that it is a function of the parameter vector $(b_1/\sigma, \dots, b_k/\sigma)$. This greatly increases the feasibility of examining the CP function.

To find the CP of this CI, we use the law of total probability.

We derive a simplified expression for this CP and show that it is a function of the parameter vector $(b_1/\sigma, \ldots, b_k/\sigma)$. This greatly increases the feasibility of examining the CP function.

Even so, finding the minimum CP over the entire space $\mathbb{R}^k$ of possible values of $(b_1/\sigma, \ldots, b_k/\sigma)$ is nearly impossible. We use a careful analysis that shows that this minimum CP can, in fact, be found by searching over a much more restricted space of values of $(b_1/\sigma, \ldots, b_k/\sigma)$.

To find the CP of this CI, we use the law of total probability.

We derive a simplified expression for this CP and show that it is a function of the parameter vector $(b_1/\sigma, \dots, b_k/\sigma)$. This greatly increases the feasibility of examining the CP function.

Even so, finding the minimum CP over the entire space $\mathbb{R}^k$ of possible values of $(b_1/\sigma, \dots, b_k/\sigma)$ is nearly impossible. We use a careful analysis that shows that this minimum CP can, in fact, be found by searching over a much more restricted space of values of $(b_1/\sigma, \dots, b_k/\sigma)$.

Also we use a new simulation method for estimating this CP, that uses variance reduction by conditioning.

Numerical illustration for data taken from Milliken & Johnson (2002)

We consider data that is taken from Chapter 3 of Milliken & Johnson (2002).

Numerical illustration for data taken from Milliken & Johnson (2002)

We consider data that is taken from Chapter 3 of Milliken & Johnson (2002).

This data set pertains to the comparison of the effectiveness of three exercise programs (treatments) on the heart rate of males with ages in the range from 28 to 35 years. A total of 24 males within this age range were chosen and 8 males were randomly assigned to each of the three treatments labelled 1,2 and 3, so that $k = 3$.

Numerical illustration for data taken from Milliken & Johnson (2002)

We consider data that is taken from Chapter 3 of Milliken & Johnson (2002).

This data set pertains to the comparison of the effectiveness of three exercise programs (treatments) on the heart rate of males with ages in the range from 28 to 35 years. A total of 24 males within this age range were chosen and 8 males were randomly assigned to each of the three treatments labelled 1,2 and 3, so that $k = 3$.

Since the aim is to compare exercise programs at a common initial resting heart rate, the initial heart rate of each of the subjects are used as a covariate.

In their illustrative analysis of this data, Milliken & Johnson (2002) begin with the one-way analysis of covariance model (1) and perform the two-stage procedure to determine the form of the model.

In their illustrative analysis of this data, Milliken & Johnson (2002) begin with the one-way analysis of covariance model (1) and perform the two-stage procedure to determine the form of the model.

$$Y_{ij} = a_i + b_i \left(x_{ij} - \bar{x} \right) + \varepsilon_{ij}$$

where
Y_{ij} is the **heart rate** of the j'th person receiving treatment i,
x_{ij} is the **initial heart rate** of the j'th person receiving treatment i,
i=1,2,3, and j=1,...,8.

We suppose that the parameter of interest θ is the difference between the expected responses of two subjects receiving treatments 1 and 2, for the same given value x^* of the covariate.

We suppose that the parameter of interest θ is the difference between the expected responses of two subjects receiving treatments 1 and 2, for the same given value x^* of the covariate.

$$\theta = \mathrm{E}(Y_1^*) - \mathrm{E}(Y_2^*) = a_1 - a_2 + (b_1 - b_2)(x^* - \bar{x})$$

where Y_1^* and Y_2^* denote the responses of two subjects receiving treatment 1 and 2, respectively, for the same value x^* of the covariate.

We suppose that the parameter of interest θ is the difference between the expected responses of two subjects receiving treatments 1 and 2, for the same given value x^* of the covariate.

$$\theta = \mathrm{E}(Y_1^*) - \mathrm{E}(Y_2^*) = a_1 - a_2 + (b_1 - b_2)\,(x^* - \bar{x})$$

where Y_1^* and Y_2^* denote the responses of two subjects receiving treatment 1 and 2, respectively, for the same value x^* of the covariate.

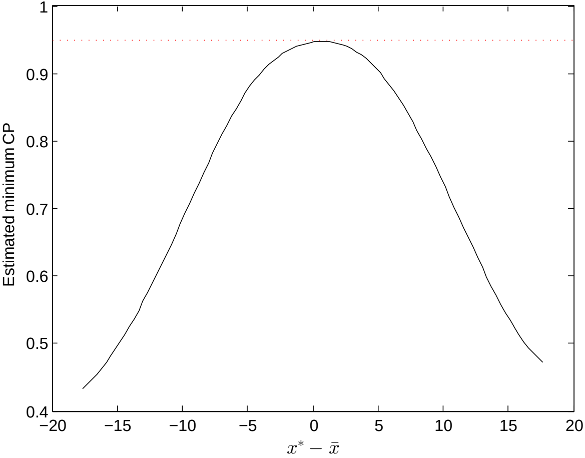
We consider the CI for θ, with nominal coverage 95%, constructed after this two-stage procedure.

For both of the F tests performed in the two-stage procedure, the significance level is chosen to be 10%.

Initially, we examine the effect of the covariate on the minimum CP of the CI for θ, where we restrict attention to values of the covariate that are within the range of values of the covariate in the data.

$$\theta = \mathrm{E}(Y_1^*) - \mathrm{E}(Y_2^*) = a_1 - a_2 + (b_1 - b_2)(x^* - \bar{x})$$

The minimum CP of the CI for θ constructed after the two-stage procedure was estimated for a grid of values of $x^* - \bar{x}$, and is depicted in the following figure.



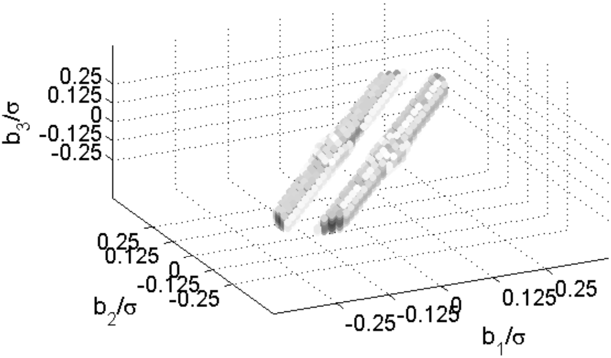
We then carry out a detailed investigation of the centrally-located parameter values for which the CP of the CI for θ is relatively small, when we fix x^* to be at its maximum distance from $\bar{x}$.

We then carry out a detailed investigation of the centrally-located parameter values for which the CP of the CI for θ is relatively small, when we fix x^* to be at its maximum distance from $\bar{x}$.

The minimum CP of the CI is estimated to be 0.4385, where this minimum CP is achieved for $(b_1/\sigma, b_2/\sigma, b_3/\sigma) \in [-0.25, 0.25]^3$.

We estimate the CPs for a grid of values of $(b_1/\sigma, b_2/\sigma, b_3/\sigma) \in [-0.25, 0.25]^3$, using $M = 10000$ simulation runs for each parameter value, and the relatively small estimated CPs (estimated CPs that are less than 0.6) are plotted using a 3-D Scatter plot as follows.

We estimate the CPs for a grid of values of $(b_1/\sigma, b_2/\sigma, b_3/\sigma) \in [-0.25, 0.25]^3$, using $M = 10000$ simulation runs for each parameter value, and the relatively small estimated CPs (estimated CPs that are less than 0.6) are plotted using a 3-D Scatter plot as follows.



Conclusion

The estimated minimum CP of the CI for θ, with nominal coverage 0.95, constructed after the two-stage model selection procedure is, to a good approximation, 0.4385.

Conclusion

The estimated minimum CP of the CI for θ, with nominal coverage 0.95, constructed after the two-stage model selection procedure is, to a good approximation, 0.4385.

In other words, the CIs for θ have minimum CP far below 0.95, showing that it is completely inadequate.

Conclusion

The estimated minimum CP of the CI for θ, with nominal coverage 0.95, constructed after the two-stage model selection procedure is, to a good approximation, 0.4385.

In other words, the CIs for θ have minimum CP far below 0.95, showing that it is completely inadequate.

Furthermore, the CP of this CI is far below 0.95 for a wide range of centrally-located values of the parameter vector $(b_1/\sigma, b_2/\sigma, b_3/\sigma)$.

References

- ABEYSEKERA, W., KABAILA, P. and YILMAZ, O. (2013) The coverage probability of Confidence Intervals in one-way Analysis of Covariance after two F tests. *Aust. N. Z. J. Stat.* **55(3) 221–234.**

- **Koehler, K. (2002) Review of** *Analysis of Messy Data, Vol. III: Analysis of Covariance* **by George A. Milliken and Dallas E. Johnson. Boca Raton, FL. Chapman and Hall/CRC. J. Amer. Statist. Assoc. 97 1206–1207.**

- **Milliken, G.A. and Johnson, D.E. (2002). Analysis of Messy Data. Volume III: Analysis of Covariance Chapman & Hall/CRC, Boca Raton, Florida.**

THANK YOU