How Students Relate Residual Behavior With Parameter Significance Tests in a Linear Model? A Case Study in a Mexican Higher Education Institution

José Luis Ángel Rodríguez Silva
Monterrey Institute of Technology, Aguascalientes, México

José Paúl Carrasco Escobar
National Polytechnic Institute, México City, México

In the implementation of a linear regression model, it is crucial to carry out significance tests to assure the proper inclusion of parameters. In addition, in order to perform such a hypothesis test, students should verify the properties of the residuals and the most important is undoubtedly the normality of the residuals. However, it is clear that there is, in general, an important gap between the hypothesis testing and residuals verification stages. This paper presents an empirical study about the most probable confusions and misunderstandings students have in relating these two mental statistical constructs in a course of Administrative Statistics in a Mexican higher education institute. One of the main results is that the order in which these two concepts are taught is essential to guarantee that there is not a disconnection between the two stages, and perhaps it would be better to first analyze the properties of the residuals and then to engage in the hypothesis testing phase.

Keywords: linear regression model, model parameters, significance tests, teaching statistics

Introduction

Regression analysis is a critical assignment to many students in professional bachelors. Indeed, several of the most important problems in marketing, finance, engineering, etc., can be solved only, in a complete sense, by using the regression analysis techniques, in order to establish a potential relationship between several variables and to predict the behavior of one (or several) of them in function to the remaining ones. In this sense, for many university programs, it is fundamental the regression analysis course and it has an important place in college statistics. As a matter of fact, regression analysis is one of the most widely applied branches of statistics.

When it is carried out a regression analysis, one of the most important facts is to show the linear model itself and to determine if it fits the data in a reasonable sense. One of the key tasks that must to be carried out is to perform statistical hypothesis tests of the estimated parameters (based on the t-Student and F distributions for the individual and joint significance, respectively) to know if such parameters should be included in the linear model. After this, by the topic sequence shown in many statistical books, it is usual to invest a great deal of time to study several aspects of the linear model, like fitness degree, the potential inclusion of some variables (to construct a multiple regression model), several model modifications, and related aspects. Until the last part
of the regression analysis module, it is studied some properties about the residual behavior, like non-normality, heteroscedasticity, autocorrelation, etc. These properties are crucial to assure the robustness of the hypothesis tests about the estimated model parameters. For example, if there is a severe lack about the normal behavior of residuals, it is very probable to conclude incorrectly about the potential inclusion of the estimated parameters. However, these residual properties are checked frequently until the last part of the regression analysis module, arising in the following two problems:

1. This temporal gap could create potential confusion and a typical student is unlikely to understand the intimate and fundamental relationship between these topics;

2. It is not infrequent not to study the residual properties due to lack of time and pressures related to cover some other parts of the course, and potentially the professor might not emphasize the intimate link between these two topics.

By the previous considerations, it seems clear that there is a need to have a deeper insight and more profound empirical results in order to perform a clearer framework in the regression analysis topic sequence and to give a coherent didactic proposal about it.

The importance about topic order has been studied by some researchers, like Moore (1997), Hoerl, Hahn, and Doganaksoy (1997), Chance and Rossman (2001), and Malone, Gabrosek, Curtiss, and Race (2010), giving some approaches from different points of view about the “correct” order topics in statistics courses. However, it is clear that in general terms, the literature is scarce and there is not enough analysis, especially empirical ones, to elucidate the complexity and possible solutions to this problem, which is critical to improve the statistics teaching-learning process.

Now, in the usual exposition of many books about regression analysis, although there is not a complete uniformity in the topic exposure, most of them have an order, something it can be named the “classic order” about it. For example, in Salvatore and Reagle (2002), it is given the next sequence:

(a) Simple regression analysis;
(b) Multiple regression analysis;
(c) Further techniques and applications in regression analysis;
(d) Problems in regression analysis;
(e) Simultaneous equation models.

It is evident that there are at least two ample academic units (“multiple regression analysis” and “further techniques and applications in regression analysis”) between the significance hypothesis tests and checking model assumptions, mostly about the residual behavior (analyzed in “problems in regression analysis”). For this reason, it is of interest to answer if this is an adequate order, in a pedagogic sense, or if it should be re-ordered such topic sequence to increase a better understanding.

With these reflections, the problem definition will be described next.

The Problem

Based on a case study in a Mexican higher education institution, the purpose of this study is to analyze which one of the following didactic sequences seems to be more appropriate in the learning-teaching process of the significance hypothesis of the linear model parameters:

Didactic Scheme 1: Classical exposure of the regression analysis models, described in textbooks like Salvatore and Reagle (2002) or Walpole, R. Myers, S. Myers, and Ye (2007), testing first the significance of
the parameters of the model, spending some time in studying intermediate topics, and then arriving it (there is enough time) to analyze the residual behavior.

Didactic Scheme 2: Joint exposure of the significance parameter tests with the residual analysis, that is, immediately after the study of the $t$-Student and $F$-Fisher tests about model parameter significance, it is analyzed the residual behavior to test the suitability of such tests.

**Method**

In the semester August-December (AD) 2012, 33 economic-administrative students (finance, marketing, and international business management) in a Mexican higher education institution were given a lecture about regression analysis, under the Scheme 1 described in the problem definition section, i.e., under a “classic order”.

In the semester January-June (JJ) 2013, 35 students with identical profile, i.e, students of finance, marketing, and international business management were given the same lecture, but this time under the Didactic Scheme 2, that is, with a joint exposure of significance test of parameters and immediately after with a complete study of residual behavior.

All students had taken a previous assignment of Administrative Statistics I with the following topics: univariate descriptive statistics (describing data sets with aid of histograms and pie graphs, and finding the usual statistics like mean, median, mode, variance, standard deviation, etc.), main random variables (discrete and continuous), and the principal properties of normal distribution (analyzing the Central Limit Theorem and the Law of Large Numbers).

In fair terms, except by the didactic sequences, all other factors remain without change. That is, it was taken the same support textbook (Salvatore & Reagle, 2002), the same informatics tool (Minitab®), in the same schedule (7:00-8:00 a.m., three times per week), and with almost the same independent academic burden (about five hours of independent work per week). Moreover, the socio-demographic and economic students’ conditions were very similar (almost the same average age, all of them are full-time students, and with a very similar mean income). By the settings, there were no confusion factors and the only significant variable changed was the didactic sequences previously described.

What it was taught was the basic theory and practice of the linear regression model. Specifically, during the classroom sessions, it was exposed the corresponding material of chapters 6 to 9 of Salvatore and Reagle (2002).

The way the lectures were given was by the problem-based learning (PBL) approach.

In order to gather information, two items were employed: (a) a quantitative approach; and (b) a qualitative-perception flavor. These two items were:

1. Supposing you have found a linear model of one dependent variable, $Y$, against two independent variables, $X_1$ and $X_2$, you are considering the linear model $Y = b_0 + b_1X_1 + b_2X_2$, and you have three $t$-Students empirical values: 3.45 for $b_0$, 3.87 for $b_1$, and 5.32 for $b_2$. The corresponding $t$-Student theoretical value is 2.83. Moreover, the empirical $F$-Fisher value is 5.21 and the corresponding theoretical value is 3.24. Finally, the Anderson-Darling $p$-value to test normality of residual is 0.001: (a) Write the corresponding sets of hypothesis tests for the model parameters; and (b) What are your conclusions about this model, i.e., do we have a correct model? Explain.
2. Given the values of the $t$-Student and $F$-Fisher statistics, as well as the Anderson-Darling $p$-values about normality of residuals, what is the relationship between these values?

In both items, students’ answers were graded in a 0-10 scale, 0 with no answer at all and 10 with a complete and satisfactory answer.

Finally, a survey was carried with several students in both groups after the application of two items in order to capture a more intuitive appreciation. Two questions were applied:

1. Do you think it is important to check the regression model assumptions?
2. Why? Please, be as specific as you can.

Results

Table 1 presents, in a compact way, the results of the two items. The confidence intervals were found with a 95% confidence level, in all instances.

Table 1

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Item 1</th>
<th>Item 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AD 2012 group</td>
<td>JJ 2013 group</td>
</tr>
<tr>
<td>Size group</td>
<td>33</td>
<td>35</td>
</tr>
<tr>
<td>Mean</td>
<td>5.38</td>
<td>7.50</td>
</tr>
<tr>
<td>Median</td>
<td>5</td>
<td>7.5</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>3.86</td>
<td>2.84</td>
</tr>
<tr>
<td>Mode</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>Modal frequency</td>
<td>9</td>
<td>15</td>
</tr>
<tr>
<td>Mean interval confidence</td>
<td>(4.01; 6.75)</td>
<td>(6.52; 8.48)</td>
</tr>
<tr>
<td>Mean difference confidence interval</td>
<td>(0.47; 3.77)</td>
<td>(0.59; 4.25)</td>
</tr>
<tr>
<td>Grades ≥ 7.0</td>
<td>16</td>
<td>24</td>
</tr>
<tr>
<td>Grade proportion ≥ 7.0</td>
<td>0.48</td>
<td>0.69</td>
</tr>
<tr>
<td>Proportion confidence interval</td>
<td>(0.31; 0.66)</td>
<td>(0.53; 0.84)</td>
</tr>
<tr>
<td>Proportion difference confidence interval</td>
<td>(-0.03; 0.43)</td>
<td>(0.03; 0.49)</td>
</tr>
<tr>
<td>Variance interval</td>
<td>(9.63; 26.05)</td>
<td>(5.29; 13.88)</td>
</tr>
<tr>
<td>Quotient variance interval</td>
<td>(0.27; 1.08)</td>
<td>(0.34; 1.38)</td>
</tr>
</tbody>
</table>

Note. Source: Data analysis from source data.

Besides, the intuitive survey results point that all the students of the JJ 2013 group, without any single exception, answered that they in fact consider very important to check model assumptions, concentrating their justification in the argument that without such validation, the model itself could be useless. Contrary, in the AD 2012 group, only 52% of students (17 of 33) were able to answer affirmatively about the importance of checking such model assumptions, and only the 33% of them (10 students) gave a reasonable justification.

Result Interpretations

The academic performance in the two items of the JJ 2013 group seems to be superior, in a 95% confidence level, with respect to the AD 2015 group, as it can be concluded by the fact that 0 does not belong to the mean difference confidence interval, although it can be noted that the grade mode was equal in both
groups (10), i.e., in both didactic sequences, the modal frequency was 10.

Not only the grade itself is interesting but the proportions of students with a 7 approved note. It is observed that in Item 1, it seems that there is no significant difference (due to the fact that 0 belongs to the confidence interval), but in Item 2, there happens a contrary fact, that is, it seems to be a significant difference.

Variances seem to be equal in both groups, because 1 is contained in both intervals.

It is conclusive the rate of students that recognizes the importance of checking model assumptions if they have the chance to link immediately study of residual behavior with the construction of the model itself. If it is not carried such connection, the students could be lost and not realize the importance of checking model assumptions, specially the normal residual behavior.

By a direct interview with several students of both groups, it seems that the joint scheme is a more reasonable choice to emphasize the robustness of significance tests with respect to residual behavior.

**Conclusion**

The didactic sequences in statistics courses are relevant, and for this reason, the choice to select a particular order can influence, in a significant way, the academic performance of students.

In this case of study research, it was proved that it is a good pedagogic choice to bind the significance tests of parameters of a linear model with the verification of the residual properties, mostly about normality. Equivalently, it does not seem to be a good strategy to allow a gap between these two topics if we wanted to maximize student understanding about this transcendental regression analysis construct.

Finally, there is a lot of work to do to find the best strategies in the topic sequences, not only in regression analysis modules, but virtually in all statistics courses. Both theoretical and empirical works are crucial in order to propose the best educative alternatives to increase student understanding.

**References**


