Validating Regression Assumptions Through Residual Plots – Can Students Do It?

Handanhal Ravinder  
Montclair State University  
Email: ravinderh@mail.montclair.edu

Mark Berenson  
Montclair State University  
Email: berensonm@mail.montclair.edu

Haiyan Su  
Montclair State University  
Email: suh@mail.montclair.edu

ABSTRACT

In linear regression an important part of the procedure is to check for the validity of the four key assumptions. Introductory courses in business statistics take a graphical approach by displaying residual plots of situations where these assumptions are violated. Yet it is not clear that students at this level can assess residual plots well enough to make judgments about the violation of the assumptions. This paper discusses the details of an experiment in progress that addresses this issue using a two-period crossover design. Preliminary results are presented.

KEYWORDS: Regression, Linearity, Homoscedasticity, Scatterplot, Cross-over design

INTRODUCTION

When assessing the assumptions of linearity, independence of error, normality, and equal spread in simple linear regression modeling, students of introductory business statistics are typically presented with an exploratory, graphical approach to the topic of residual analysis (for example, Donnelly, 2015: Anderson, Sweeney & Williams, 2016; Levine, Szabat, & Stephan, 2016). Except for the assumption of independence of error where the Durbin-Watson technique (1951) is presented, no basic text covers any confirmatory approach to any of the other assumptions.

Berenson (2013) however, raises a basic question – are inexperienced undergraduate business students really capable of critically assessing residual plots? If this is not the case, it would be imperative to introduce confirmatory, inferential approaches to supplement the graphical, exploratory approaches currently found in texts and taught in classrooms.

This study is a first step toward answering this question. It focuses on how well students can validate the assumptions of linearity and homoscedasticity through residual plots. The other regression assumptions will be examined in a follow-up study.
VALIDATING REGRESSION ASSUMPTIONS

LITERATURE REVIEW

Anscombe (1973), in a seminal paper on the graphical approach to residual analysis, clearly demonstrated the importance of data visualization for enhancing an understanding of what a data set is conveying and for assisting in the model-building process in simple linear regression. The scatterplot, however, has been a particular focus of research, given its popularity in the representation of bivariate data, and much work has been done on how well humans assess correlations from scatterplots (Pollock, 1960; Bobko & Karren, 1979; Lauer & Post, 1989; Meyer, Taib, & Flascher, 2005; Doherty, Anderson, Angott, & Klopfer 2007). Cleveland, Diaconis, & McGill (1982) studied the impact of display factors (such as size of the plotting character, the overall size of the display, the orientation and the size of the point cloud within the frame) on how well subjects were able to assess the underlying correlations. Lauer & Post (1989) demonstrated that density of scatterplots (number of points in the plot) influences the estimation of correlation. Cook and Weisberg (1994) clearly showed why the selected aspect ratio of a scatterplot is essential to its understanding. Berenson (2013), concerned that a graphical residual analysis of the very important homoscedasticity assumption in a regression model may, in itself be insufficient, suggested that such an exploratory approach to residual analysis be supplemented with a confirmatory approach and recommended White’s test (1980) for this purpose.

DEVELOPING THE EXPERIMENT

Three treatment variables were selected for this experiment – aspect ratio, plot density, and plot characteristic.

Aspect Ratio: Per Cook and Weisburg (1994), the aspect ratio of a chart is clearly essential to its understanding so aspect ratio became the primary treatment factor in the study. Two levels were selected, (i) the aesthetic 1.61 golden ratio of height to width (or 0.62 height to width in landscape chart format) and (ii) an aspect ratio of twice that value.

Plot Density: Plot density, i.e., the number of plotted points observed in a chart has an impact on how well subjects assess the information in a scatterplot (Lauer & Post, 1989). Again two levels were selected – 20 points per plot versus 60 points per plot.

Plot Characteristic: Plot characteristic has to do with the particular regression assumption that is complied with or violated. There is also the potential for confounding departures from linearity with departures from homoscedasticity. Four types of plot characteristics were selected – one indicating no violation of linearity or homoscedasticity, one possessing indications of significant departure from linearity without evidence of departure of homoscedasticity, one demonstrating significant increases in variability without evidence of departure from linearity, and, lastly, one displaying evidence of significant departure from both linearity and homoscedasticity.

To achieve the desired objective an experiment was then devised to evaluate the percent of correct student assessments to a series of residual plots generated by using the two levels of aspect ratio, with the two levels of plot density and the four types of residual plots nested within each aspect ratio. A two-period crossover design model was employed (Jones & Kenward, 1989; Sheskin, 2011). For each group it was determined that samples of 168 student subjects would be needed to achieve a power of 0.90 to detect a 5 point percent difference (i.e., the effect size) in correct response to the two aspect ratio treatments when testing at the 0.05 level of significance (Pocock, 1983; Fleiss, 1986; Cohen, 1988; Grissom & Kim, 2012).
IRB approval was obtained for the study. Subjects were students (with majors primarily in business) in various introductory undergraduate statistics courses taught by various instructors. They were offered extra credit in their respective statistics classes for their participation in the experiment. The experiment was spread over two semesters. The fall semester constituted group 1 and the spring semester, group 2. Participants were directed to an online survey where they were shown a series of 18 residual plots, 16 that were part of the study and two that were provided for measuring interval validity. Participants were asked to respond to two questions for each residual plot, one focusing on the presence or absence of linearity in the plot and the other on the presence or absence of homoscedasticity in the plot. A randomized sequence of eight residual plots was developed by combining the two levels of plot density with the four levels of plot characteristics. A separate plot used for internal validity was placed in the fifth position. In the fall semester this sequence of nine residual plots was presented to the group 1 subjects with aspect ratio 1 in period 1 followed by aspect ratio 2 in period 2. There was no break between period 1 and period 2. In the spring semester, the group 2 subjects received the same sequence of plots, except that the order of the aspect ratios was switched. The experimental scheme is shown below where “P” is a “study plot” and “C” is a control plot.

<table>
<thead>
<tr>
<th>Group 1: Fall</th>
<th>Group 2: Spring</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>P2</td>
</tr>
<tr>
<td>P9</td>
<td>P10</td>
</tr>
</tbody>
</table>

The first of the two internal validity residual plots (C1) displayed clearly obvious violations to both assumptions while the other (C2) clearly displayed no violation to either of the assumptions. Students who did not answer each of these four ‘obvious’ questions correctly were excluded from the study analysis. Either they did not understand the assessment task correctly or did not take the task seriously enough.

**INTENDED PRIMARY AND SUPPLEMENTARY RESEARCH QUESTIONS**

- **Primary Research Question 1**: How good are students at identifying the different residual plots correctly? This question pertains to the main purpose of this paper and will have important implications for the teach of introductory business statistics course as well as for textbooks.
- **Primary Research Question 2**: Does a chart’s aspect ratio affect visual perception? This research question is expected to shed light on the importance of aspect ratio selection in student ability to correctly graphically assess the linearity and homoscedasticity assumptions through graphic residual analysis.

Given the results of this inferential analysis several additional research questions will need to be addressed:

- **Supplementary Research Question 1**: How good are subjects at identifying the linearity versus the homoscedasticity assumptions in the residual plots?
- **Supplementary Research Question 2**: Controlling for aspect ratio, how does plot density affect perception ability?
• **Supplementary Research Question 3**: Controlling for plot density, how does aspect ratio affect perception ability?

• **Supplementary Research Question 4**: With respect to the various questions is there evidence that subject graphic perception is superior to random guessing?

### THE TWO-PERIOD CROSSOVER DESIGN MODEL

A two-period crossover design model was employed in this study. The main advantage to this design is that it is efficient — it uses lots of information through a crossover and is statistically a more powerful design approach than an alternative parallel group model approach which would treat two independent groups separately. The main disadvantage to the two-period crossover design is that the subjects in each group need to be exposed to a sequence of 18 (16 measurable and 2 internal validity control) residual plots and make two responses on each — unfortunately providing more opportunity for boredom or distraction to confound the results than would be had with the alternative parallel group model approach.

Excluding the internal validity control question responses, each student responds twice to each of the residual plots presented — one response involving the student’s perception of the presence or absence of pattern (as opposed to random noise) which concerns violation of the linearity assumption and the other response involving the student’s perception of the presence or absence of increasing vertical spread which concerns violation of the homoscedasticity assumption.

This two-period crossover design results in 32 binary responses per student subject, each simply scored as correct (1) or incorrect (0). The total number of correct responses to questions pertaining to one of two particular aspect ratios (i.e., the treatment conditions) will then be transformed to percent correct response and reported as a score on a 0 to 100 scale.

### Table 1. Data Layout for a Two-Period Crossover Design.

<table>
<thead>
<tr>
<th>Group 1</th>
<th>Group 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subject</td>
<td>Pd. 1</td>
</tr>
<tr>
<td>1</td>
<td>$X_{111}$</td>
</tr>
<tr>
<td>2</td>
<td>$X_{112}$</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>n</td>
<td>$X_{1n}$</td>
</tr>
</tbody>
</table>

Provided that the assumptions of the classical, parametric tests are met, Chassan (1964), Grizzle (1965), and Hills and Armitage (1979) suggested four test procedures to be considered for a two-period crossover design based on the pooled t-test statistic. These test procedures are.
displayed below. On the other hand, under more general conditions, Koch (1972) proposed a corresponding set of four test procedures based on the Mann-Whitney-Wilcoxon rank statistic. Regardless of whether the parametric or distribution-free set of tests is employed, the fourth test procedure is equivalent to what would be used for assessing direct treatment effects in the alternative parallel group design.

The following symbols are used to represent the various effects to be studied:

- $\lambda$ represents the residual carryover effects of aspect ratio from one period to another and evaluated in Test 1 of a two-period crossover design. If such residual carryover effects are observed, Test 4 must be used in lieu of Test 3 for investigating direct treatment effects.
- $\pi$ represents the period effects evaluated in Test 2 of a two-period crossover design. If such period effects are observed it would indicate that boredom or lack of concentration likely occurred in period 2 and would require the use of Test 4 instead of Test 3 for investigating direct treatment effects.
- $\tau$ represents the direct treatment effect evaluated in Test 3 of a two-period crossover design. A direct treatment effect is the effect that a treatment (here aspect ratio) has during the period it is administered.
- $\tau| (\lambda \text{ or } \pi)$ represents the direct treatment effect evaluated in Test 4. Because of carryover or period effects, the second period assessments from each group are discarded and the two period-1 assessments (made under different aspect ratios) are compared.

### Four Classical Test Procedures

Each of the four test procedures is based on the pooled $t$-test. In general, let the pooled $t$-test statistic be represented by

$$t_y = \frac{\bar{Y}_{1.} - \bar{Y}_{2.}}{\sqrt{\frac{S^2_{P_1}}{n} + \frac{S^2_{P_2}}{m}} \frac{S^2}{nm}}$$

where $S^2_{P_1} = \frac{\sum_{i=1}^{n}(Y_{1i} - \bar{Y}_{1.})^2 + \sum_{i=1}^{m}(Y_{2i} - \bar{Y}_{2.})^2}{(n-1) + (m-1)}$.

Using a level of significance $\alpha$, the null hypothesis is rejected if $t_y < t_{(n+m-2), \alpha/2}$ or if $t_y > t_{(n+m-2), 1-(\alpha/2)}$.

**Test 1: Testing for Equality of Residual Carryover Effects**

For Test 1 the null hypothesis is $H_0 : \lambda_1 = \lambda_2$ and the test statistic is $t_y = t_{\lambda}$. Here $\bar{Y}_{1.} = \bar{T}_{1.}$, the mean of the totals of each subject’s response to both treatment A and treatment B in group 1, and $\bar{Y}_{2.} = \bar{T}_{2.}$, the mean of the totals of each subject’s response to both treatment B and treatment A in group 2. Moreover, $S^2_{P_1} = S^2_{P_2}$, the pooled variance for these data.

**Test 2: Testing for Equality of Period Effects**

Assuming $\lambda_1 = \lambda_2$, for Test 2 the null hypothesis is: $H_0 : \pi_1 = \pi_2$ and the test statistic is $t_y = t_{\pi}$. Here $\bar{C}_{1.} = \bar{C}_{1.}$, the mean of the differences or crossover changes between each group 1 subject’s response to treatment A in the first period and treatment B in the second period, and $\bar{C}_{2.} = \bar{C}_{2.}$, the mean of the differences or crossover changes between each group 2 subject’s
response to treatment A in the second period and treatment B in the first period. Moreover, $S_{\lambda_1}^2 = S_{\lambda_2}^2$, the pooled variance for these data.

**Test 3: Testing for Equality of Direct Treatment Effects**

Assuming $\lambda_1 = \lambda_2$, for Test 3 the null hypothesis is: $H_0 : \tau_1 = \tau_2$ and the test statistic is $t_\tau = t_\tau$. Here $\bar{Y}_{1-} = \bar{D}_{1-}$, the mean of the differences between each group 1 subject’s response to treatment A in the first period and treatment B in the second period, and $\bar{Y}_{2-} = \bar{D}_{2-}$, the mean of the differences between each group 2 subject’s response to treatment B in the first period and treatment A in the second period. Moreover, $S_{\lambda_1}^2 = S_{\lambda_2}^2$, the pooled variance for these data. (Note that $\bar{D}_1$ here is identical to $\bar{C}_1$ in Test 2).

**Test 4: Testing for Equality of Direct Treatment Effects**

If $\lambda_1 \neq \lambda_2$ and significant carryover effects are found (or, if $\pi_1 \neq \pi_2$ and significant period effects are found), for Test 4 the null hypothesis is: $H_0 : \tau_1 = \tau_2$ and the test statistic is $t_\tau = t_{\pi_1\pi_2\tau}$. Here $\bar{Y}_{1-} = \bar{X}_{1-}$, the mean of each group 1 subject’s response only to treatment A in the first period, and $\bar{Y}_{2-} = \bar{X}_{2-}$, the mean of each group 2 subject’s response only to treatment B in the first period. Moreover, $S_{\lambda_1}^2 = S_{\lambda_2}^2$, the pooled variance for these data. Note that this test procedure, making use of only the first period responses in groups 1 and 2, is equivalent to the test for equality of direct treatment effects in a parallel group design.

If the assumption of normality is not met, the distribution-free analogs of each of these tests (Koch, 1972) will be employed.

**PRELIMINARY RESULTS**

A total of 262 students initially participated in the fall semester. Based on their responses to the four internal validity questions, 185 group 1 students were retained for the study’s primary and supplementary analyses while 78 were excluded. Thus, the fall semester student subject participation rate was 70.2%. Data collection for the spring semester (group 2) is currently in progress.

Ravinder, Berenson, & Su (2016) reported preliminary findings based on a descriptive analysis of the responses from group 1, the fall semester student subjects. The first primary research question was addressed preliminarily and summarized below. The second primary research question, however, is inferentially-based and requires data from both the fall and spring semester student subject groups.

To identify a residual plot correctly, subjects would have to answer both the question on linearity and the question on homoscedasticity correctly. As displayed in Figure 1, subjects in the fall semester had the fewest errors of classification when there were no violations of either kind – linearity or homoscedasticity.
Proper assessments were made 84.6% of the time. Nevertheless, when there was at least a violation in one of the assumptions, proper identification of the charts were made only between 55.1% and 62.3% of the time, calling into question whether textbook and classroom instruction using only graphic residual analysis approaches are sufficient.

Once the data are obtained from group 2 (the spring semester) a similar descriptive analysis will be made individually and then by combining the two groups. The findings here are expected to be an important part of the overall study.

**CURRENT STATUS OF RESEARCH**

The experiment is in the final stages of obtaining assessments from the spring semester student subjects (group 2). Spring course section enrollments suggest that number of participants should be similar to the fall semester. Having both groups’ responses will enable inferential conclusions to be drawn pertaining to the role of aspect ratio as an important factor influencing the ability of students to correctly identify the linearity and homoscedasticity assumptions in residual plots and supplementary analyses will provide important information concerning graphic literacy based on plot density and plot characteristic. Interesting and pedagogically useful results are anticipated.

**REFERENCES**


