Regression vs. AOV: Which to Choose?

By: Josh Sheinberg

Abstract:

In this study, we explore the differences between two common statistical methods (Regression and Analysis of Variance) on predicting the average household adjusted income across the 50 states in 2015. These two methods will be compared in two different settings, each using two predictor variables; one with a significant interaction present between the two predictors and one without a significant interaction. These methods are compared from the context of the research question being considered, the statistical results, the graphical results, and the resultant answer (interpretation of the statistical and graphical results). In the end, we find that neither model is objectively better than the other. We do find, however, that the added complexity of Regression models does not always result in an answer that differs from the simpler AOV model when there is no interaction present. With an interaction, we find that AOV does not always tell the whole story.

Introduction:

All sciences use some sort of statistical analysis. However, it is not always clear what model is best for the situation. With this, there is a strong importance of educating researchers on the pros and cons of the models they have the option of using. Details of these analyses are

often overlooked, and can be critical to an understanding of the results produced. In this paper, some of the details of both Regression and AOV will be presented.

When explaining the economic wellbeing of a state, there are many distinct ways to measure it. In this paper we will use income adjusted for cost of living. Although this does not fully explain the economic wellbeing of a particular state, it allows for an even comparison across the US. The question of interest is "Can we predict adjusted income across the states?". To predict adjusted income among the states, there are many different variables to choose from. For the purposes of this study, I gathered macroeconomic and educational data from 2015 in each of the states, and chose four, as presented in Table 1.

Table 1
Variable definitions and descriptive statistics

Variable	Definition	Mean	Min	Median	Max
Income ($)	Household income adjusted for cost of living (US Bureau of Economic Analysis)	57,568	47,092	56,897	69,456
Federal Funding ($)	Federal funding per capita for public schools (National Center for Education Statistics)	116.04	78.89	106.35	360.39
Taxes ($)	Average tax rate: $\frac{High\ Tax\ Bracket + Low\ Tax\ Bracket}{2}$ (Federation of Tax Administration)	3.58	0.00	3.75	7.60
Unemployment (%)	Unemployment Rate (Bureau of Labor Statistics)	5.27	2.70	5.4	7.50
Teacher Quality	Proportion of core academic classes taught by teachers who are considered highly qualified: bachelor's degree, state certification, and competency in the subject taught. (National Center for Education Statistics)	0.96	0.73	0.97	1

Notes. All data are from 2015 and measures for all 50 states, with Washington D.C. added. Parenthesis in the definition mark the source the data were gathered from. "<1" indicates a fraction censored to one.

Methods:

Conceptually, any univariate prediction model uses one or more variables to predict a single response variable. In a model with a single predictor, the generic question of interest is:

> "Can the information contained in the predictor variable be used in some systematic fashion (model) to explain the variation we see in the response variable?"

If the answer is yes, then we can predict values of the response based on our knowledge of the values of the predictor. For instance, if we have knowledge that there is an inverse relationship (model) between state level unemployment and adjusted income, then we would know that we could increase adjusted income per capita within our state (response) if we could lower unemployment (predictor). Such single predictor models are simple and easy to explain, but unfortunately are not realistic. Most models today are more complicated. More complicated can be defined in a variety of ways, but in this paper it will be limited to an expansion of our model to include multiple predictor variables. Fortunately, this is also the definition used by most researchers.

In a model with two predictors, the generic question of interest becomes;

> "Can the information contained in the first predictor variable, the information contained in the second predictor variable, and the information contained in the combined influence of the first and second predictor variables be used in some systematic fashion (model) to explain the variation we see in the response variable?"

It can be seen that this generic question contains three sub-questions; (1) is the generic single predictor question associated with the first predictor, (2) is the generic single predictor question associated with the second predictor, and (3) is a new question involving the simultaneous

influence of both predictors. In the language of statistics, the single predictor question would be called a main effect and this new question would be called the joint or interactive influence. The addition of more predictors to the model would add greater complexity, more variables, to the question; however, it would not add greater analytic complexity. The model would still be composed of main effects (associated with a single predictor) and interaction effects (associated with any combination of predictors). Since the primary emphasis of this paper is the comparison of analytic methods, I have limited the models to their most simplistic level, two predictor variables (two main effects terms and a single interactive term).

In statistics there are several methods that can be used to answer the two predictor variable generic question. The two most popular of these are multiple Regression and two-way Analysis of Variance (AOV). Given the ratio nature of the data presented in Table 1, the statistical method best suited to answer the generic question is Multiple Regression. However, the definition of the interactive influence of the two predictor variables within the Regression context is highly problematic, not universally accepted, analytically complex, and interpretationally difficult. This creates such a problem in Regression that most professional researchers resolve the associated interaction problems by simply excluding the interactive sub-question from inclusion in the generic question. In essence, they have resolved to assume that the two predictors do not have an interactive influence on the response variable. This particular decision results in the analytic problem known as multicollinearity, which I will not address in this paper.

Table 2
Binary variable definitions

Variable	Definition	Low (below median)	High (above median)
Income ($)	Household income adjusted for cost of living (US Bureau of Economic Analysis)	0	1
Federal Funding ($)	Federal funding per capita for public schools (National Center for Education Statistics)	0	1
Taxes ($)	Average tax rate: $\frac{High\ Tax\ Bracket + Low\ Tax\ Bracket}{2}$ (Federation of Tax Administration)	0	1
Unemployment (%)	Unemployment Rate (Bureau of Labor Statistics)	0	1
Teacher Quality	Ratio of core academic classes taught by teachers who are considered highly qualified: bachelor's degree, state certification, and competency in the subject taught. (National Center for Education Statistics)	0	1

Notes. All data are from 2015 and measures for all 50 states, with Washington D.C. added. Parenthesis in the definition mark the source the data were gathered from. "<1" indicates a fraction censored to one.

In contrast, many researchers are unwilling to make this rather bold assumption of no interactive influence which results in a severe limitation in the answer to the generic question. As a consequence, they have sought out alternative methods which can be easily expanded to incorporate the interactive term in the model. The most common model in which the interaction term is included is Analysis of Variance. Unfortunately, while this model includes all of 3 of the sub-questions in the two predictor situation, it cannot accommodate the predictors as defined in Table 1. Within the context of Analysis of Variance, the predictors must be defined in a categorical manner. For instance, it would be possible to take the information for the predictors presented in Table 1 to redefine them as presented in Table 2 based on a median adjustment. In Table 2, each variable is now defined as being below (score = 0) or above (score = 1) the median. Hence, a state with an unemployment rate below 5.4 (the median unemployment rate as presented in Table 1) would receive a score of 0 using the variable definition in Table 2. Such scoring for all the variables in Table 2 produces categorical predictor variables suitable for

considering Analysis of Variance models. Even though the full two predictor generic question can now be addressed, it comes with a potentially serious caveat, which is "How much information contained in the variables as defined in Table 1 have been lost through their redefinition in Table 2?" This is not an easy question to answer, except in specific problem contexts, which is one of the goals of this paper.

In summary, in research we are commonly and realistically interested in complex questions involving two or more predictors. However, we are confronted with no ideal statistical method of answering the full question. Although Multiple Regression is fully capable of using the information at its highest level (Table 1), it does so at the expense of sacrificing the sub-answer arising from the interactive term. In contrast, although Analysis of Variance is fully capable of addressing all of the sub-questions, it does so at the expense of sacrificing some of the information contained in the original variables. Which of the two approaches is best is a highly debated question among professional researchers and even statisticians. It is not the goal of this paper to answer this question, but rather consider solutions from both perspectives and to provide guidance on their comparison which might be able to lead to a specific answer in a specific situation.

In general, Regression models are more analytically complex and lend themselves to more complicated interpretations (answers to our questions) by accessing the higher level information contained in the predictors. In contrast, Analysis of Variance models are more analytically simple and easily lend themselves to simpler interpretations, which can be easily communicated to others in presentations and papers. This leads us to the logical question, "Does

greater complexity and more complicated interpretations actually result in answers that differ from the simpler ones?" Unlike the question in the paragraph above, this question is not often considered by anyone. It is a goal of this paper to attempt to answer this question in the context of the two specific analyses that will be conducted.

The first specific situation considered the impact on the analysis methods and on the interpretational results when the interactive component of the model does not exist. [In essence, the situation in which the major assumption of Regression is satisfied.] The second situation considered the impact on the analysis methods and on the interpretation results when the interactive component of the model does exist.

Results Part 1:

Question 1: How do unemployment rate and teacher quality effect the adjusted income of the states in 2015?

In this situation the predictor variables will be unemployment rate (ur) and teacher quality (tq), and the response variable will be adjusted income (ai).

Multiple Regression

The interaction term used in the multiple regression was the simple multiplicative influence, where $interaction = unemployment\ rate * teacher\ quality$. Hence, the regression model associated with the full question of interest is;

$$predicted\ (ai) = b_0 + b_1 * ur + b_2 * tq + b_3 * (ur * tq) \quad (1)$$

where b_0, b_1, b_2, and b_3 are the associated regression coefficients.

The t-test result associated with the interaction effect (ur*tq) in (1) above was not significant ($t(47) = 0.04$, $p = 0.97$). As a consequence, the interaction term was pooled into the error term and produced the classic multiple regression model with only main effect terms, as seen in (2).

$$predicted\ (ai) = b_0 + b_1 * ur + b_2 * tq \quad (2)$$

The statistical tests and regression coefficients associated with (2) appear in Table 5 of the Appendix, and result in;

$$predicted\ (ai) = 105{,}610 - 2{,}348 * ur - 37{,}163 * tq \quad (3)$$

The classic graphic depiction of a simple Regression result portrays the single predictor variable in the horizontal axis and the single response variable in the vertical axis. It would be possible to extend this classic graph into three dimensions (two for the two predictors and one for the response). Unfortunately, such a graph is difficult to appropriately convey in a two-dimensional medium such as in this report. However, it is possible to employ an alternative method which can be displayed in two dimensions. It capitalizes on the simplicity of the Regression graph by presenting the first predictor in the horizontal axis and providing separate regression lines for a select value of the second predictor. The selected values that I used for the second predictor are five values depicting the full range of responses at discrete points. These are at two standard deviations below the mean, one standard deviation below the mean, the mean, one standard deviation above the mean, and two standard deviations above the mean.

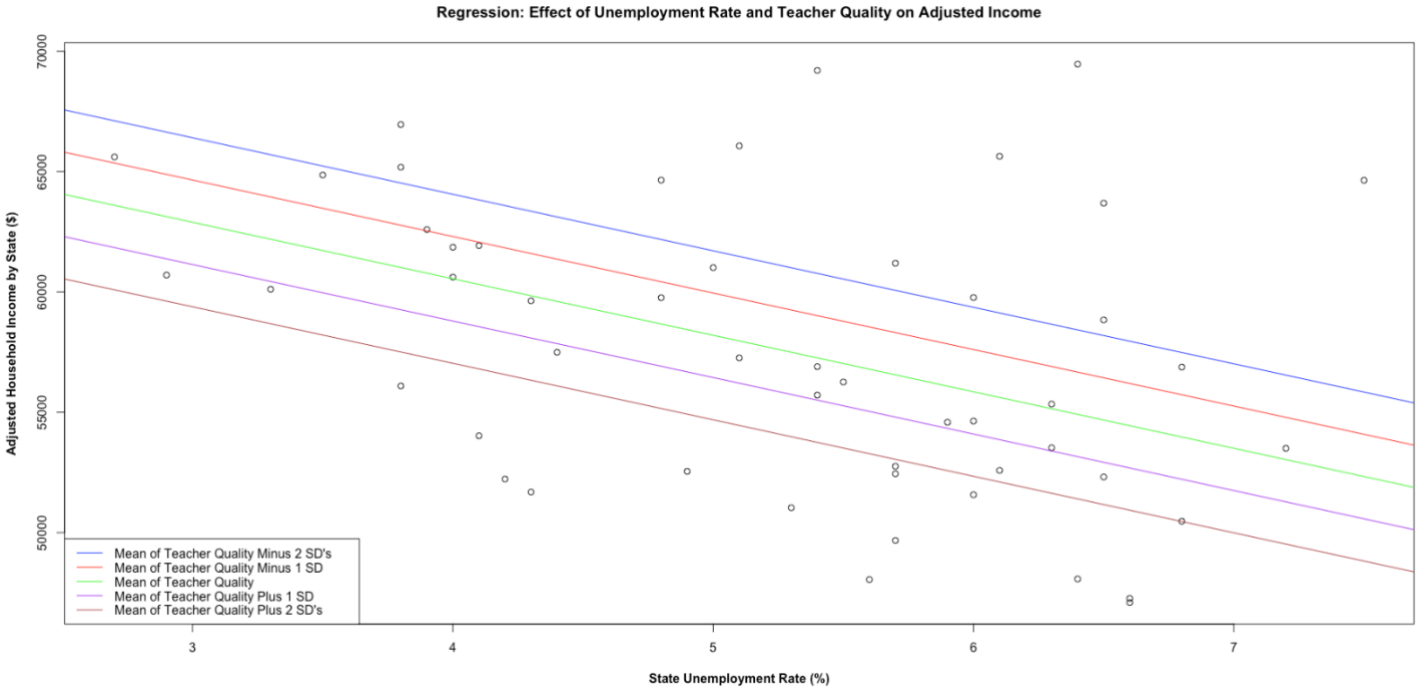
Using this method, the regression associated with (3) is calculated for the five values of Teacher Quality in Table 3 which are graphically presented in Figure 1 (the first predictor is unemployment rate and the second predictor is teacher quality). As can be seen in Table 3 each

of the five regression equations produced are identical in their slopes and only differ in the constants with each equation being separated by the influence of Teach Quality (b_2*tq). To those familiar with regression, it is not surprising that the graphical results presented in Figure 1 reflect parallel regression lines.

Table 3: Regression Equations Associated with Specific Values of Teacher Quality (s = .047, m=.96)

Teacher Quality	Equation
$-2s = (-2) * (.047) = -.094$ $m - 2s = .96 - .094 = .866$	$b_0 + b_1 * ur + b_2 * tq = 105610 - 2348 * ur + (-37163) * (.866) = 73426.8 - 2348 * ur$
$-1s = (-1) * (.047) = -.047$ $m - 1s = .96 - .047 = .913$	$b_0 + b_1 * ur + b_2 * tq = 105610 - 2348 * ur + (-37163) * (.913) = 71680.1 - 2348 * ur$
$0 = (0) * (.047) = 0$ $m - 0 = .96 - 0 = .96$	$b_0 + b_1 * ur + b_2 * tq = 105610 - 2348 * ur + (-37163) * (.96) = 69933.5 - 2348 * ur$
$1s = (1) * (.047) = .047$ $m + 1s = .96 + .047 = 1.01$	$b_0 + b_1 * ur + b_2 * tq = 105610 - 2348 * ur + (-37163) * (1.01) = 68075.4 - 2348 * ur$
$2s = (2) * (.047) = .094$ $m + 2s = .96 + .094 = 1.05$	$b_0 + b_1 * ur + b_2 * tq = 105610 - 2348 * ur + (-37163) * (1.05) = 66588.9 - 2348 * ur$

Figure 1
Regression model with no interaction



In Figure 1, the relationship between unemployment rate and income is shown by each of the regression lines of Table 3. Each differently colored line represents a different level of teacher quality. It is clear from this figure that the basic regression relationship between Unemployment Rate and Adjust Income is negative; as unemployment rate increases adjusted income correspondingly decreases. Specifically, I found that for each 1 percent increase in the unemployment rate adjusted income decreases by $2,348. This relationship is the same regardless of the level of Teacher Quality since each regression based on the different levels of Teach Quality has the same regression slope (Table 3). Hence, the overall interpretation of Figure 1 is that, for the lowest level of teacher quality (blue line) adjusted income declines as the unemployment rate increases. This interpretation is identical for each of the lines. The difference between the lines reflects the differences between levels of teacher quality. I found that as teacher quality increases, the lines appear progressively lower in the figure. Thus, the impact of

an increase in teacher quality on the regression is to reduce the starting point of the regression (left side of the graph), but does not influence the relationship between unemployment rate and adjusted income. This is the direct result of the insignificant interaction term.

In economics terms, this negative relationship between unemployment and income can be explained by a rising unemployment slowing the economy. If there is a downturn in the economy, it is common for unemployment to rise. With this downturn, adjusted average household income would decrease.

AOV

The full Two-Way Analysis of Variance model associated with our 2 predictor question is

$$predicted\ (ai) = \ \bar{a\imath} + a_i + \ b_j + \ c_{ij} \qquad (4)$$

where $\bar{a\imath}$ = mean of adjusted income,
a_i = main effects associated with unemployment rate
b_j = main effects associated with teacher quality
c_{ij} = interaction effects associated with unemployment rate and teach quality
For i = 0 for low unemployment (below median) and 1 for high unemployment (above the median) and
j = 0 for low teacher quality and 1 for high teacher quality

In essence, converting teacher quality to a binary variable (two values; 0 and 1) is the ultimate reduction of teacher quality. In Table 3 and Figure 1, teacher quality was reduced from its original level of measurement in proportion to five distinct points (-2s, -1s, 0, +1s, +2s). Rather than having five distinct values of teacher quality, I reduced the variable to only two values (low and high). The resulting F-test associated with the interaction effects was not significant ($f(1,47) = 0.95$, $p = 0.33$). Due to the insignificance of this term, the interaction was pooled into the error term. As a consequence, (4) reduces to

$$predicted\ (ai) = \bar{a}\imath + a_i + b_j\ . \qquad (5)$$

This produced an AOV model with only the main effects and no interaction. The test statistics associated with (5) appear in Table 6 of the Appendix.

The summary of the Analysis of Variance results appear in Table 4 and can be used to produce Figure 2.

Table 4: Adjusted Income Means Associated with the Analysis of Variance Model

Unemployment Rate	Teacher Quality	Adjusted Income
Low	Low	$60,411
Low	High	$59,299
High	Low	$56,264
High	High	$55,152

Figure 2
AOV model with no interaction

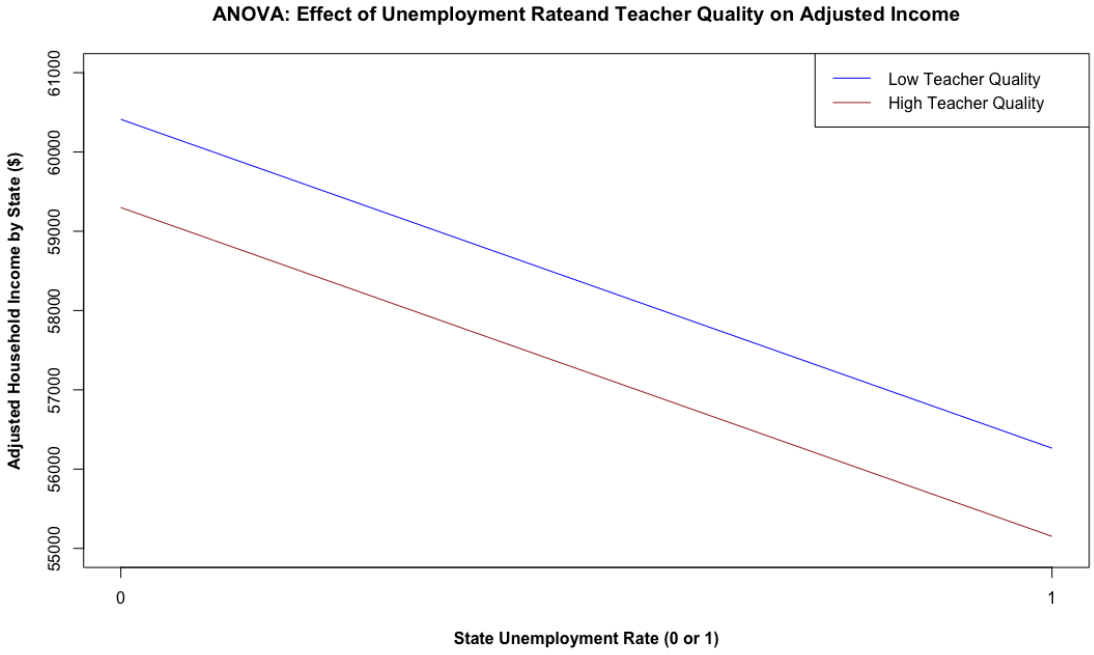


Figure 2 shows the relationship between unemployment (low versus high) and income taking into account teacher quality; low teacher quality (blue line) and high teacher quality (brown line). Each of the means in Table 4 are represented by the end points of each line.

As with the regression, the lack of an interaction term has resulted in a graph in which the lines are parallel. In fact, there are many similarities between Figures 1 and 2. In each the lines slope from the upper left hand corner of the graph to the lower right hand corner. Although it would be inappropriate to interpret the lines in Figure 2 as slopes, they do in fact have extremely similar interpretations. In Figure 2, the interpretation of both lines is that as unemployment rate increases (goes from low to high) adjusted income decreases (has a low mean value). This appears to be exactly the same interpretation as was expressed for Figure 1 using regression. The only difference between Figures 1 and 2 is that unemployment rate is measured on a continuum in Figure 1 and only at the end of this continuum in Figure 2. In addition, it is seen that the same relationship exists between unemployment rate and adjusted income for all values of teacher quality (parallel lines). It is once again the case that lower teacher quality is universally associated with higher adjusted income (the blue line is above the brown line). The only difference between Figures 1 and 2 is that in Figure 1, teacher quality has five progressively larger values of quality and in Figure 2 there are only two.

Comparison

For these data in Part 1, the statistical results (non-significant interaction), the figures, and the answers to the original question of multiple regression and Analysis of Variance appear quite similar. One could easily ask the question, "With so many similarities are these two methods as different as many statisticians would lead us to believe?".

In part it may appear as if I stacked the deck in favor of the two methods being seen as similar, through the selection of the variables that I used in the prediction, since the functional relationship between each of the predictor variables and the response variable was linear. As a consequence, any two values lying on the continuum between low and high on one of the predictors had to be in a straight line (linear) relationship as seen in Figure 1. With were only two selected values from the predictor (low and high), they would have to express exactly the same relationship, which is what I have shown in Figure 2, since two points is the definition of a straight line. Hence, the linear nature of the data forced the two figures to look the same and should not come as a surprise. Although this could be perceived as a serious negative in this paper, it is by far the assumed nature of variable relationships throughout many disciplines. However, many disciplines, the biological sciences and in economics in particular, the relationships between the variables are in fact non-linear. Even though these data were linear, the possibility of non-linear data reveals the major difference between these two methods.

Regression is not simply the expression of the relationship between predictor variables and a response variable, but it is the functional expression of this relationship. This regression has several benefits over Analysis of Variance. First, since the predictor is measured on a continuum, it is possible to anticipate (predictor) what might happen in the response variable between sampled values of the predictor. This is impossible within the classic Analysis of Variance context in which the variables are seen to exist only at and the results can only be interpreted at the sampled points. This is a major limitation of Analysis of Variance. However, as I showed above, if the data are linear, then this limitation is so minor as to simply vanish. This is decidedly not the case when the data are non-linear. Regression can make use of the pattern expressed across the variety of sampled points to establish a functional relationship and allow intermediate

prediction. This could only be done in Analysis of Variance if the researcher had the clairvoyance to be able to sample the exact values of functional change prior to collection of the data. Even in well researched areas this is a near impossibility. As a result, in a situation in which it is not known what functional relationship might exist between the predictors and response variables, Regression is always the superior choice. However, in a situation such as the one presented here in Part 1, Analysis of Variance is not only an easier to apply alternative, it is also a much easier to interpret with literally no loss of meaningful information due to the categorization of the predictor variables. Most people will find the two lines of Figure 2 to be easier to understand than the five lines of Figure 1.

Results Part 2:

Question 2: How do average tax rates and federal funding to public schools jointly affect the adjusted income of the states 2015?

In this situation, the predictor variables will be tax rate (tr) and federal funding to public schools (ps), and the response variable will be adjusted income (ai).

Multiple Regression

Once again, the interaction term used in the multiple regression was the simple multiplicative influence. In this situation interaction = tax rate * federal funding to public schools. Hence, the regression model associated with the full question of interest is

$$predicted\ (ai) = b_0 + b_1 * tr + b_2 * ps + b_3 * (tr * ps) \qquad (4)$$

where b_0, b_1, b_2, and b_3 are the associated regression coefficients.

The t-test result associated with the interaction effect (tr*ps) in (4) above is statically significant (t(47) = -3.72, p-value = 0.0005). The statistical tests and regression coefficients associated with (4) appear in Table 9, and result in

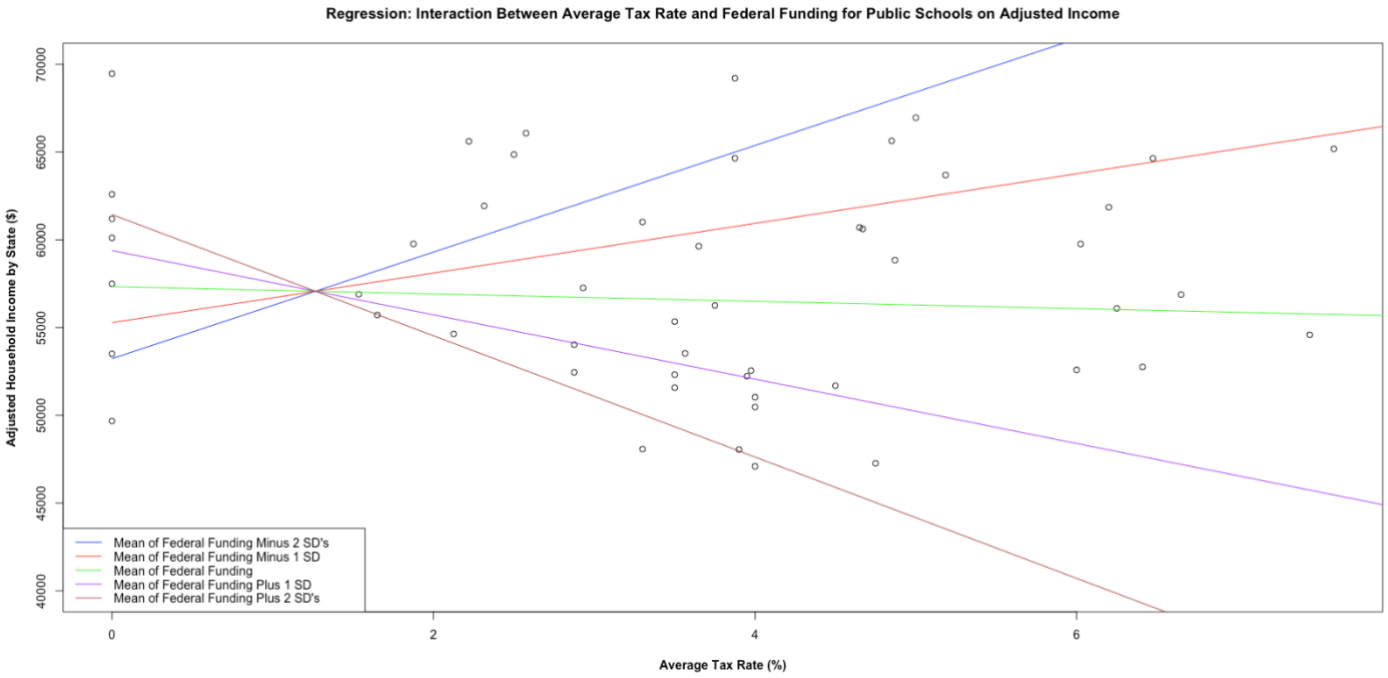
$$predicted\ (ai) = 51{,}793 + 4169 * tr + 47.7 * ps - 37.7 * (tr * ps) \qquad (5)$$

Similar to the procedure in Part 1, the regression associated with (5) is calculated for the five values of Federal Funding in Table 7 which are graphically presented in Figure 3 (the first predictor is average tax rate and the second predictor is federal funding). As seen in Table 7, each of the 5 regression equations produced have unique slopes. The regression results presented in Table 7 generate Figure 3.

Table 7: Regression Equations Associated with Specific Values of Federal Funding (s = 43, m=116)

Teacher Quality	Equation
$-2s = (-2) * (43) = -86$ $m - 2s = 116 - 86 = 30$	$b_0 + b_1 * tr + b_2 * ps + b_3 * tr * ps = 51793 + 4169 * tr + (47.7) * (30) - (37.7) * tr * (30) = 53224 + 3038 * tr$
$-1s = (-1) * (43) = -43$ $m - 1s = 116 - 43 = 73$	$b_0 + b_1 * tr + b_2 * ps + b_3 * tr * ps = 51793 + 4169 * tr + (47.7) * (73) - (37.7) * tr * (73) = 55275 + 1417 * tr$
$0 = (0) * (43) = 0$ $m - 0 = 116 - 0 = 116$	$b_0 + b_1 * tr + b_2 * ps + b_3 * tr * ps = 51793 + 4169 * tr + (47.7) * (116) - (37.7) * tr * (116) = 57326 - 204 * tr$
$1s = (1) * (43) = 43$ $m + 1s = 116 + 43 = 159$	$b_0 + b_1 * tr + b_2 * ps + b_3 * tr * ps = 51793 + 4169 * tr + (47.7) * (159) - (37.7) * tr * (159) = 59377 - 1825 * tr$
$2s = (2) * (43) = 86$ $m + 2s = 116 + 86 = 202$	$b_0 + b_1 * tr + b_2 * ps + b_3 * tr * ps = 51793 + 4169 * tr + (47.7) * (202) - (37.7) * tr * (202) = 61428 - 3446 * tr$

Figure 3
Regression model with an interaction



In Figure 3, the relationship between average tax rate and income is shown by each of the regression lines of Table 7. Each differently colored line represents a different level of federal funding. This figure shows that the relationship between taxes and income relies on the level at which federal funding is at. Because of this, the relationship is sometimes positive, and other times negative. Table 7 shows this by having both positive and negative coefficients for tax rate at different levels of federal funding. Specifically, when is federal funding is low (blue line), I found that for each 1 percent increase in the tax rate, income increases by $3,038. Looking at the other extreme would be when federal funding is at a high level (brown line), a 1 percent increase in tax level results in a $3,446 *decrease* in income. The other levels of federal funding have their own unique slopes, illustrated by the other lines in Figure 3. The relationship between federal

funding and income is similarly dependent on the tax rate. According to Figure 3, if the tax rate is below about 1.5%, then there is positive correlation between funding and income. This is shown by going from the blue line to lines which represent higher levels of taxes. However, once taxes are above this 1.5% mark, the relationship turns to a negative correlation. On the right side of the graph, going from the blue line to the other higher levels of federal funding, income decreases. This is a direct result of the significant interaction term.

Because of this significant interaction, there is a need for both of the predictor variables. With just one or another, the relationship with the dependent variable would not be accurate. For example, if this regression only had tax rate without federal funding, then the relationship could be positive or negative. This would result in only a partial understanding between the relationship of income and taxes. When federal funding is added, though, the relationship starts to be more complete. By having the second predictor variable, it becomes possible to better understand the relationship with income. The significant interaction demonstrates the need for both predictor variables.

AOV

Moving onto an AOV, the predictor variables of tax rate and federal funding have again been reduced to two values, below the median (0) and above the median (1). The F-test associated with the interaction was significant ($F(1,47) = 6.74$, $p = .013$). Due to this, the general equation is

$$predicted\ (ai) = \bar{a}\imath + a_i + b_j + c_{ij} \qquad (6)$$

This produced an AOV with the interaction term, as well as the main effects. The test statistics associated with (4) appear in Table 10 of the Appendix.

The summary of the Analysis of Variance results appear in Table 8 and can be used to generate Figure 4.

Table 8: Adjusted Income Means Associated with the Interaction Analysis of Variance Model

Average Tax Rate	Federal Funding	Adjusted Income
Low	Low	$57,266
Low	High	$58,198
High	Low	$60,681
High	High	$53,476

Figure 4
AOV model with an interaction

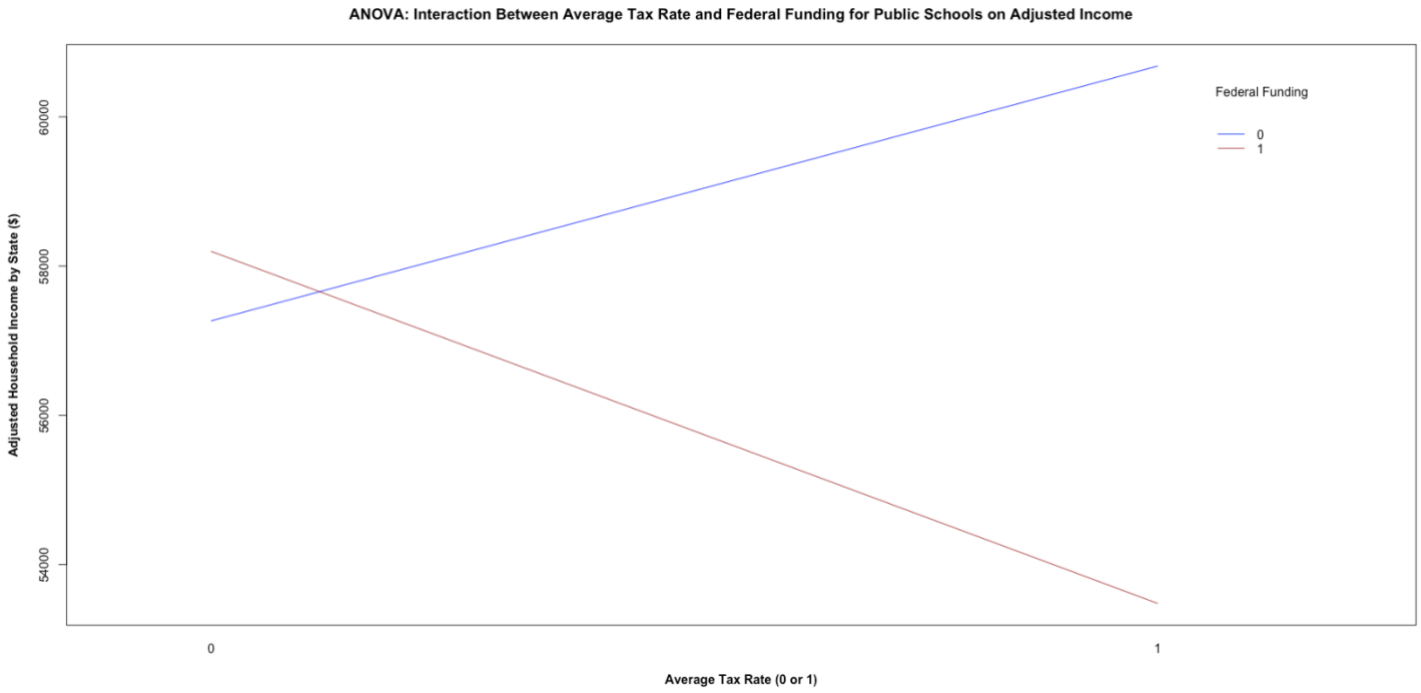


Figure 4 shows the relationship between taxes (low versus high) and income, taking into account federal funding; low federal funding (blue line) and high teacher quality (brown line). Each of the means in Table 8 are represented by the end points of each line.

As with the regression in Figure 3, the lines cross on the left side of the graph. Along with this, Figure 3 has other similarities with Figure 4. Specifically, I found two basic relationships; one in a positive context (blue line), and one in the negative context (brown line). The blue line in each figure starts in the middle of the graph on the left side, then slopes up to the upper right-hand corner. Similarly, the brown line starts close to the same place, and slopes down to the bottom left. Although it is inappropriate to interpret slopes on an AOV model, the interpretations of these two lines are very close to the regression model. In Figure 4, the interpretation of the of the blue line (low federal funding) is that as tax rates increase (goes from low to high), income also increases (has a high mean value). In contrast, the brown line (high federal funding) shows that as taxes increase, income decreases. This appears to have a very similar interpretation to Figure 3.

Comparison

For the data in Part 2, the statistical results (significant interaction), the figures, and the answers to the original question of multiple regression and Analysis of Variance appear considerable similar.

Regression has a few distinct benefits over Analysis of Variance through this example. The main benefit of this method is the specificity of an interaction term. Unlike AOV, it is possible to see what happens at each level of a predictor variable as the other predictor changes. In AOV, only the low and high of each situation is expressed. This can be observed by the green line in Figure 3, which shows almost no change, since it is nearly flat. This relationship is entirely missing from the AOV, since we only see the extremes and loses the medial points.

The problem with Regression in this context, though, is that the interaction term is not unique. This makes it exceptionally difficult to explain why the interaction selected is the

correct one. The simplest possible interaction between variables is to multiply the variables with each other. With having ratio predictor variables, there are an infinite amount of ways to manipulate them, such as creating ratios by dividing one variable by another or even using higher order forms of the variables, such as squaring, inverting, square-rooting, etc.… Having so many ways to shape each of the predictor variables, there becomes an infinite possibility of interactions. In contrast, AOV only has one unique interaction. Having categorical predictor variables generates only the base interaction of $x_1 * x_2$ as an option. This is because it is not possible to manipulate categorical variables in the same way as ratio variables (exponents, trigonometric functions, etc…). Because Regression does not have a unique interaction, justifying why the interaction that is in the model is the correct one can be difficult. In all, the AOV is simpler, but not complete because of the missing medial values.

Conclusion:

Although there is no objectively better model to use, there are pros and cons with both Regression and AOV. When comparing the respective models, the statistical results, figures, and the answers to the original questions all appear to be quite similar. The regression, though, has many benefits over the AOV. With Regression having a predictor variable that is on a continuum (ratio), it allows for predictions in-between sampled values. This is a limitation of AOV, since it can be desirable to anticipate what happens between the sampled values of the predictors. However, since the data is linear, this limitation becomes very minor. When there is no significant interaction in the model, the added complexity of regression is not worth it over an AOV.

In comparison to Part 2 of the results, though, an interaction can change the way the models function. With these models, there was a distinct difference: the green line in Figure 3. This relationship, when the containment variable is at a medial point, is not shown at all in the AOV. When there is a significant interaction in the model, the added complexity of Regression is necessary to explain the relationships in full over an AOV.

In all, as shown in these two cases, the added complexity of regression models does not always result in an answer that differs from the simpler ANOVA model. As shown in this paper, with no interaction, AOV can tell the same story with less complexity. With a significant interaction, though, the added complexity of a Regression can be justified through the medial values that are lost in an AOV. This conclusion is counterintuitive from what these models are built to have accomplished. Regression is not contracted for interactions, while AOV is assembled specifically to allow for interactions. What we found, though, is the opposite. Although AOV is made for interaction terms, it doesn't always convey the full relationship between the predictors and dependent variables. On the other side, regression (assuming linearity) does not have any distinct advantages over AOV, which is much simpler. With this, I encourage researchers to more often compare these models before going with one or another.

Appendix

Table 5

Regression results predicting Income with Unemployment and Teacher Quality

Variable	t(48)	p-value	Coefficient
Unemployment	-3.36	0.002	-2,348
Teacher Quality	-2.15	0.037	-37,163
Overall F(2,48)= 6.26	p-value = 0.0038	R^2 = 0.21	

Table 6

AOV results predicting Income with Unemployment and Teacher Quality

Variable	F(1,48)	p-value
Unemployment	5.95	0.018
Teacher Quality	0.46	0.500
Overall F(2,48)= 3.21	p-value = 0.049	R^2 = 0.12

Table 9

Regression results predicting Income with Tax Rate and Federal Funding

Variable	t(47)	p-value	Coefficient
Tax Rate	3.51	0.001	4169.1
Federal Funding	2.33	0.024	47.74
Tax Rate* Federal Funding	10.15	0.0005	-37.37
Overall F(3,47)= 4.88	p-value = 0.005	R^2 = 0.24	

Table 10

AOV results predicting Income with Tax Rate and Federal Funding

Variable	F(1,47)	p-value
Tax Rate	0.08	0.783
Federal Funding	4.26	0.045
Tax Rate* Federal Funding	6.74	0.013
Overall F(3,47)= 3.69	p-value = 0.018	R^2 = 0.19