Wert, Mark (DEP)

From:	Richard Joy [RJoy@sierraresearch.com]
Sent:	Thursday, June 28, 2001 4:58 AM
To:	Mark. Wert@state. ma. us (E-mail)
Subject:	RE: GD Im240 vs. MA99 All data comparisons

Mark, we are really trying to work with George, but when he sends something like the attached email it really upsets me ... and I am not even the one it is directed at here at Sierra. There is no way that Garrett ever told him that we believed a high order polynomial made engineering sense, and his sarcastic comment about Garrett having taken three statistics courses is really a slap in the face. He is beginning to act more and more unprofessional with us, particularly Garrett, and we do not deserve to be treated this way.

I read George's email just now after working into the wee hours here on another project, so I am probably less tolerant than I might be if I was well rested. However, George's behavior seems to be getting worse not better. If you read Garrett's email to him (which I read and approved before it was sent), you'll see that it is pretty inocuous and actually is in agreement with George's view that a linear regression model makes the most sense. He obviously doesn't agree with the decision to use the clipped data, but I believe he is off-base on his views on this issue. Part of his message reads:

> ...at this time we possibly have tens of thousands of vehicles failing tests in MA because the cutpoints were set up incorrectly, without the necessary adjustment. Artificial narrowing the range will only diminish relevance and validity of the equations, thus increasing the number of false passes. Although I agree that nonlinearity in the upper part of the range is irrelevant from the point of view of the test decision making, it is not easy to say where in the range the equations becomes so "nonlinear" that the points beyond the limit become unimportant.

While we have not analyzed this issue yet, I doubt that anywhere near tens of thousands of vehicles are falsely failing in MA due to the use of NY-based conversion factors. His comment about false passes also makes little sense. While the selection of 2 x the cutpoint as where to clip the data to produce interim conversion factors is relatively arbitrary, it provides a comfortable safety margin around the pass/fail points. I know George believes we need a lot more data in the mid-high emission ranges in order to develop regressions with acceptable accuracy. I agree completely that this would be ideal; however, simply saying that we cannot make any judgments or produce useful analysis results until we collect these data does not reflect the reality of the situation (i.e., how hard it is to find/test such high emitters and the limited time we have left for the AZ study). George needs to offer constructive comments, not negative and increasingly personal ones. Garrett hasn't seen George's message yet, but I can imagine how he is going to feel when he reads it.

-----Original Message-----
From: Zeliger, George (DEP) [mailto:George.Zeliger@state.ma.us]
Sent: Wednesday, June 27, 2001 8:43 PM
To: Garrett Torgerson
Cc: Mark. Wert@state. ma. us (E-mail); Richard Joy; Michael St. Denis
Subject: RE: GD Im240 vs. MA99 All data comparisons

Garrett,

Another file -- this time based on real data -- that might be of interest to you. According to your directions, I took the 91 to 95 Cars HC data and truncated them by removing points (actually, just one point) with the MA99 value exceeding two times the April 2001 cutpoint (1.2x2=2.4). I ran then several regressions, including those high degree polynomials you seemingly were so interested in, as well as the exponential curve; by some reasons unknown to me, Excel excluded the logarithmic fit from the very beginning (maybe, it is smarter than I think of it).

As you can see from those plots that Excel has generated, the behavior of the curves becomes more and more erratic as the degree grows. Beginning with the third degree, the curve is not convex; fourth and higher degree curves are not even monotone. I am extremely curious about how you will explain a non-monotone relationship between the two kinds of measurements.

To the right of the graph you'll find 16 columns with coefficients of the polynomials (the degrees are shown in bold italics on top), followed by the values of the Adjusted Squared Correlation Coefficient as well as the Average Squared Residuals for each equation. Neither of the values is a valid characteristics to base the choice of the polynomial on; however, many not so experienced users of statistics would use them as such, so I included the numbers. Based on the minimal value of the ASR and the maximal value of the ASCC (both highlighted blue), the 13th degree polynomial should be chosen -- an obviously absurd conclusion. To better understand its absurdity just look at the behavior of the polynomial coefficients as the degree grows -- their absolute values grow unlimited, while their signs alternate -- this alone would cause so large computational errors in an attempt to convert a MA99 value to an IM240 one, that the final result would have not a single correct digit.

Besides, the behavior of the ASR and ASCC are not monotone, which again suggests that the whole endeavor is meaningless -- some very particular features of the sample used to generate the equations as well as -- to even greater extent -- computational errors accumulating in the course of calculating the values of the coefficients and all the statistical characteristics totally overweight that small amount of useful information hidden in the data.

Finally, as I mentioned, neither ASCC nor ASR can be used for a scientifically sound choice of the best estimate because of the problems associated with the statistical independence -- or, rather, lack of it -- between parameters of different equations (polynomial or others) calculated on the same sample. Use of orthogonal polynomials could help a little bit in reducing computational errors as well as in struggling with the lack of independence -- but only just a little. Unfortunately, Excel -- a statistically mediocre software, containing many well known elementary errors -- does not allow that use. Besides, in case of truly nonlinear relationships like the exponential function (you certainly remember that polynomials are still linear in their parameters) all those orthogonality issues are irrelevant and worthless.

I am sure I have not told you anything new -- you knew everything from those three courses in statistics you have taken.

It shouldn't be concluded from what I said so far that I only except the simple linear (i.e., linear both in parameters and variables) model no matter what -- same way as you do not intend to use "a non-linear HC regression at all costs." What I am trying to explain is that lousy data cannot be improved by using computationally complex methods that might look impressive to a naive layman. The general principle I am referring to is that it is impossible to bake a cherry pie from a pile of ... . By no means this is my discovery -- I am sure your professors whose background you know better than mine have told you about that.

I would only like to make some applications to our situation. Particularly, I don't believe it is a good idea to truncate the collected data. First, as you can see from the comparison of the "truncated" and original equations on the plot, although truncation does change the equation and its characteristics, the final result is still not so usable -- the new equation is no better than similar equations for CO and NOx, and the problem of the cutpoint conversion still remains open.

Second, it s a very bad practice in general to remove some data points -- because we don't like them -- post hoc rather than analyze them on the spot. For example, the inspector in AZ should have noticed that abnormal 45th value that I removed at your suggestion, and should have made a qualified decision on whether there was something wrong with the vehicle, or the conditions of the test, etc. Since both MA99 and GD stations showed unusually high values, I would conclude that the point actually does belong in the range of possible values -- in other words, it is possible to expect vehicles that will generate values between the "two times the cutpoint" and the eliminated point to show up sooner or later. The truncation leads to underepresentation of those vehicles, and thus to the distortion of the relationship. It is incorrect to say that
"use of this clipped data set improves the accuracy of the regression where it is most important -- in the range of the pass/fail cutpoints."

Extrapolation of the equation built only on the data points from the neighborhood of an arbitrarily chosen cutpoint will only lead to wrong decisions of both kinds in the future. As I mentioned already, at this time we possibly have tens of thousands of vehicles failing tests in MA because the cutpoints were set up incorrectly, without the necessary adjustment. Artificial narrowing the range will only diminish relevance and validity of the equations, thus increasing the number of false passes. Although I agree that nonlinearity in the upper part of the range is irrelevant from the point of view of the test decision making, it is not easy to say where in the range the equations becomes so "nonlinear" that the points beyond the limit become unimportant.

Besides, I would like to remind you that one of the most important goals of the AZ study is to estimate the relationship between the two kinds of measurements at large, since this is necessary for the estimation of the reduced excessive emission. Without that goal in mind, solely for the purpose of finding the binary "pass-fail" decision making cutpoint, we don't need regressions at all. I told about that long ago, during one of the first Michael's visits to Boston, maybe as long as a year ago.

George

-----Original Message-----
From: Garrett Torgerson [mailto:GTorgerson@sierraresearch.com]
Sent: Wednesday, June 27, 2001 1:37 PM
To: 'Zeliger, George (DEP)'
Cc: Mark. Wert@state. ma. us (E-mail); Richard Joy; Michael St. Denis
Subject: RE: GD Im240 vs. MA99 All data comparisons

George,

Thanks for your regressions. Regarding the non-linear HC regression, from what I can tell, you are right in that the evidence at this time does indicate that a linear model will work. You are also right that the quality of the data plays a factor in making the decision. In fact, the evidence at this time indicates that a number of different regression models will work. Once the dataset grows, however, it might be easier to determine which regression model is most appropriate.

Until that time, however, we do not want to rule out the consideration that a non-linear regression may work best. Based upon what we know about FID and NDIR benches, we do not expect the relationship between the two HC benches to be linear. We believe that this will be more apparent for dirtier vehicles which are under-represented in the dataset you have at this time. Hopefully they will not be as underrepresented in the final dataset.

I don't mean to give the impression that we intend to use a non-linear HC regression at all costs. If the final data indicate that a linear model is most appropriate, that is fine. As an interim measure, we have recommended to Mark that revised linear conversion factors for all three pollutants developed from a "clipped" set of the AZ correlation data (limited to MA31 values no more than 2 x the highest cutpoint) be used to replace the conversion factors currently included in the software. Use of this clipped data set improves the accuracy of the regression where it is most important -- in the range of the pass/fail cutpoints - while also addressing the concern that the expected nonlinearity in the higher HC ranges could bias the regression results if a linear model were to be used. It may be that this approach will continue to make the most sense even after we obtain the final data set from AZ. Hopefully, the final data set will include more scores in the higher emissions ranges and allow a better evaluation of this issue.

Garrett
<< File: HC Trunc Cars 91-95 .xls >>