Rev. 5/82
BU-760-M

Leverage and Regression Through the Origin

GEORGE CASELLA*

The problem of deciding whether an intercept or no-intercept model is more appropriate for a given set of data is a problem with no simple solution. Often, the underlying physical situation will suggest an appropriate model; however, there still may be interest in assessing which model best fits the data or is the better predictor. In this article a different interpretation of regression through the origin is derived, that of a full fit to the original data set augmented by one further point. Examination of the leverage and influence of the augmented data point can provide help in comparing the models.

KEY WORDS: Leverage points; No-intercept model.

1. INTRODUCTION

In simple linear regression, unless there is information to the contrary, the equation $y = \alpha + \beta x + \epsilon$ is often fitted to the data. However, there are many situations where it is reasonable to constrain the line to pass through the origin. Such a constraint will usually arise from the physical characteristics of the variables measured, and in these situations the no-intercept model $y = \beta_0 x + \delta$ may be more appropriate.

* George Casella is an Assistant Professor, Biometrics Unit, Cornell University, Ithaca, NY 14853. This research was supported by National Science Foundation Grant No. MCS81-02541 and is BU-760-M in the Biometrics Unit Series. The author thanks the Associate Editor and referees for their many helpful comments, which led to a much improved version of this paper.

If the primary concern of the experimenter is to fit the data as well as possible, or to obtain a good prediction equation, he may be faced with the task of choosing between the no-intercept and intercept model. Of course, it may be the case that a more complicated model than either of the two considered here is appropriate, but here we will only be concerned with straight line relationships.

Hahn (1977) discusses various difficulties encountered when fitting models with no intercept, and notes that many of the usual statistics (such as R^2 and the model F) are not comparable between the intercept and no-intercept models. One of his suggestions is to base comparisons on residual standard deviations, which are comparable. Marquardt and Snee (1974) are primarily concerned with mixture models. They show that even though mixture models can be written in a form without an intercept, conventional "no-intercept" statistics are inappropriate, and can lead to erroneous conclusions. More recently, Gordon (1981) argues that the way R^2 is usually calculated for the no-intercept model is misleading, leading to an over-estimation of the adequacy of the fit. This is because the total sum of squares for the no-intercept model is not corrected for the mean and, since it is always greater than the corrected sum of squares, the no-intercept R^2 is usually higher than the R^2 for the intercept model. Gordon suggests using the corrected sum of squares when calculating R^2 for the no-intercept model, in order to provide a better basis for comparing the models. Thus, the problem of evaluating the no-intercept versus intercept model is not an elementary one, and some thought must be given to develop an appropriate basis of comparison.

In this paper a new way of interpreting regression through the origin is introduced; one in terms of leverage points. (For a good introduction to the concept of leverage in regression see Hoaglin and Welsch (1978).) This new interpretation can be of use in understanding the difference between the full fit and the fit forced through the origin, and may possibly be employed as either a diagnostic tool or a teaching aid. It is shown that regression through the origin is equivalent to fitting the full model to a new data set. This new data set is composed of the original observations plus one additional point that is derived from the original observations. Evaluation of the leverage possessed by this new point is equivalent to evaluating whether $\alpha = 0$ in the full model. However, the leverage approach seems slightly more flexible, and provides a graphical aid which might facilitate the task of deciding between the models.

2. LEVERAGE POINTS AND REGRESSION THROUGH THE ORIGIN

2.1 Augmenting the Data Set

Let (x_1,y_1), (x_2,y_2), $\cdots$, (x_n,y_n) be n data points observed according to

$$y_i = \alpha + \beta x_i + \epsilon_i \quad , \tag{2.1}$$

where the experimental errors, ϵ_i, are independently normally distributed with mean 0 and variance σ^2 . The least squares estimates for α and β are

$$\hat{\beta} = \frac{\sum_{i=1}^{n} (x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n} (x_i-\bar{x})^2} , \qquad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} \quad , \tag{2.2}$$

where $\bar{x} = (1/n)\sum_{i=1}^{n} x_i$, $\bar{y} = (1/n)\sum_{i=1}^{n} y_i$. If the regression is forced through the origin, then it is assumed that the data are observed according to

$$y_i = \beta_0 x_i + \delta_i \quad ,$$

and the least squares estimate of β_0 in this model is

$$\hat{\beta}_0 = \frac{\sum\limits_{i=1}^{n} x_i y_i}{\sum\limits_{i=1}^{n} x_i^2} \quad . \tag{2.3}$$

If the original data set is augmented with a new observation $(x_{n+1}, y_{n+1}) = (n^*\bar{x}, n^*\bar{y})$, where $n^* = n/[(n+1)^{\frac{1}{2}} - 1]$, then fitting the full model to the augmented data set is equivalent to forcing the original regression through the origin. This follows from the easily verified identities

$$\begin{gathered} \sum_{i=1}^{n+1} (x_i - \bar{x}_{n+1})(y_k - \bar{y}_{n+1}) = \sum_{i=1}^{n} x_i y_i \quad , \\ \sum_{i=1}^{n+1} (x_i - \bar{x}_{n+1})^2 = \sum_{i=1}^{n} x_i^2, \ \sum_{i=1}^{n+1} (y_i - \bar{y}_{n+1})^2 = \sum_{i=1}^{n} y_i^2 \quad , \end{gathered} \tag{2.4}$$

where

$$\bar{x}_{n+1} = [1/(n+1)] \sum_{i=1}^{n+1} x_i, \quad \bar{y}_{n+1} = [1/(n+1)] \sum_{i=1}^{n+1} y_i \quad .$$

The position of the point $(n^*\bar{x}, n^*\bar{y})$, relative to the other points, determines whether the new point has high or low leverage. The leverage of the new point can be used to decide if the regression through the origin is more appropriate than the model which includes an intercept term.

2.2 Assessing the Leverage of a Data Point

The leverage, h_{ij}, of a data point y_j is the amount of influence that data point has on each fitted value $\hat{y}_i$.

A predicted value, say $\hat{y}_i$, can be written as

$$\hat{y}_i = \sum_{j=1}^{n} h_{ij} y_j \quad , \tag{2.5}$$

where

$$h_{ij} = \frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{\sum_{k=1}^{n} (x_k - \bar{x})^2} \quad . \tag{2.6}$$

The h_{ij} show how each observation y_j affects the predicted value y_i . More importantly, however, h_{ii} shows how y_i affects $\hat{y}_i$, and is quite useful in the detection of influential points. The relative size of h_{ii} can give us information on the potential influence y_i has on the fit.

For purposes of comparison, it is fortunate that the values h_{ii} have a built-in scale. The matrix $H = [h_{ij}]$ is a projection matrix, and from the properties of projection matrices it can be verified that $0 \le h_{ii} \le 1$ and $\sum_{i=1}^{n} h_{ii} = p$, where p is the number of coefficients to be estimated. Thus, on the average, h_{ii} should be approximately equal to p/n . Hoaglin and Welsch suggest, as a rough guideline, paying attention to any data point having $h_{ii} > 2p/n$.

The h_{ij} values depend only on the experimental design (the x's), and not on the results of the experiment (the y's), hence a data point with high leverage may not necessarily have an adverse effect on the fit. One measure of the effect of data point i on the fit is the Studentized residual, r_i^* . This is the standardized residual corresponding to y_i when (x_i, y_i) has been omitted from the fit. Using the subscript "(i)" to denote that data point i has not been used in an estimate, we have

$$r_i^* = \frac{y_i - (\hat{\alpha}_{(i)} + \hat{\beta}_{(i)} x_i)}{[\widehat{\text{Var}}(\hat{\alpha}_{(i)} + \hat{\beta}_{(i)} x_i)]^{\frac{1}{2}}} \quad . \tag{2.7}$$

$\widehat{\text{Var}}(\hat{\alpha}_{(i)} + \hat{\beta}_{(i)} x_i)$ is the estimated variance of $\hat{\alpha}_{(i)} + \hat{\beta}_{(i)} x_i$, and a little algebra will show

$$\widehat{\text{Var}}(\hat{\alpha}_{(i)} + \hat{\beta}_{(i)} x_i) = \hat{\sigma}^2_{(i)} \left[1 + \frac{\sum_{j \neq i} (x_j - x_i)^2}{(n-1) \sum_{j \neq i} (x_j - \bar{x}_{(i)})^2} \right] , \qquad (2.8)$$

where $\hat{\sigma}^2_{(i)}$ is the residual mean square from the "not-i" fit. A large value of r_i^* suggests that data point i has a large impact on the fit. Since r_i^* follows a t distribution with $n - p - 1$ degrees of freedom, the significance of any particular r_i^* can be assessed.

2.3 The Leverage of the Augmented Point

We are concerned here with the influence of the $(n+1)$st data point $(n^*\bar{x}, n^*\bar{y})$. Using (2.6), (2.7), and (2.8) it is straightforward to calculate

$$h_{n+1} = \frac{1}{n+1} \left[1 + \frac{n^2 \bar{x}^2}{\sum_{i=1}^{n} x_i^2} \right] \quad \text{and} \quad r^*_{n+1} = \frac{\hat{\alpha}}{\hat{\sigma} \left[1 + \frac{\bar{x}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right]^{\frac{1}{2}}} , \qquad (2.9)$$

where $\hat{\alpha}$ and $\hat{\sigma}^2$ are, respectively, the estimated intercept and residual mean square from the full fit on the original n data points. As can be seen, r^*_{n+1} is exactly the t statistic that tests $H_0 : \alpha = 0$. Thus, the impact of $(n^*\bar{x}, n^*\bar{y})$ on the fit concerns only the intercept. A large value of r^*_{n+1} indicates that the two regressions (with and without $(n^*\bar{x}, n^*\bar{y})$) will have very different coefficients, and a decision must be made whether or not to include $(n^*\bar{x}, n^*\bar{y})$ in the fit. This interpretation seems more flexible than the classical one of accepting the full fit if r^*_{n+1} is large.

There is still another interpretation of r^*_{n+1} which can be useful in evaluating the no-intercept models. As noted before, Hahn (1977) recommends comparing the intercept versus no-intercept models on the basis of their residual errors. Some algebra will verify that

$$(r^*_{n+1})^2 = (n-1)\,\frac{\hat{\sigma}^2_0}{\hat{\sigma}^2} - (n-2) \quad , \tag{2.10}$$

where $\hat{\sigma}^2_0$ and $\hat{\sigma}^2$ denote, respectively, the residual mean squares from the regression through the origin and the full regression. The statistic $(r^*_{n+1})^2$ measures exactly what is recommended by Hahn, the relationship between residual variances. Notice, also, that this interpretation of r^*_{n+1} is consistent with that of (2.9); small values of $(r^*_{n+1})^2$, which support H_0: $\alpha = 0$, also support H_0: $\sigma^2_0 \leq \sigma^2$. Moreover, it is straightforward to check that $(r^*_{n+1})^2$ is equal to the C_p statistic (Mallows, 1973) for assessing the adequacy of the no-intercept model as a subset of the intercept model.

These results can be generalized to include the case where the regression line is forced through some arbitrary point (x_0, y_0) . If we define new variables x_i' and y_i' by $x_i' = x_i - x_0$ and $y_i' = y_i - y_0$, then forcing the original data through (x_0, y_0) is equivalent to forcing the adjusted data through (0,0) . There is also a straightforward generalization to multiple linear regression; the original data is augmented with the point $(n^*\bar{x}, n^*\bar{y})$, where $\bar{x}$ is a vector containing the means of the independent variables. The full fit to the augmented data set is still equivalent to fitting the no-intercept model to the original data.

The impact of the augmented data point on the fit becomes clearer when h_{n+1} is also examined. It is, perhaps, more instructive to write h_{n+1} in the equivalent form

$$h_{n+1} = \frac{1}{n+1}\left[1 + n\left(\frac{\bar{x}^2}{\sigma^2_x + \bar{x}^2}\right)\right] \quad , \tag{2.11}$$

where $\sigma^2_x = (1/n) \sum_{i=1}^{n} (x_i - \bar{x})^2$. Thus, the impact of the augmented data point increases with $(\bar{x}/\sigma_x)^2$, and we can expect the greatest discrepancy between the full fit and the fit through the origin when $\bar{x}$ is large compared to σ_x . In such cases, the augmented data set will seem to be composed of two distinct clusters; one composed of the original data and one composed of $(n^*\bar{x}, n^*\bar{y})$. This information can be of some help in cases where the primary concern is finding the best linear prediction equation. Since the augmented point will be at some distance from the original data, a plot of the augmented data set will give a more "global" view. If straight line extrapolation is realistic, assessing whether the augmented point could, in fact, be a physically valid data point would help in deciding if it is reasonable to force the regression through the origin. Indeed, it may be possible to take an observation near the augmented x value, which may help to confirm linearity and extend the range of applicability of the fitted equation.

The distance between the augmented data point and the original data is proportional to $n^{\frac{1}{2}}$. Thus, as n increases, we expect the augmented point to move further away from the original data. An explanation of this fact is provided by examining the leverage of the augmented point relative to the original data set. When the full fit is performed on the augmented data set, two coefficients are estimated, so $\sum_{i=1}^{n+1} h_{ii} = 2$. Thus, $2 - h_{n+1}$ provides a measure of the leverage of the original n data points, and a little algebra shows that the leverage of the augmented point relative to the original data is

$$\frac{h_{n+1}}{2-h_{n+1}} = \frac{\frac{1}{n}\left\{1 + [(n^*-1)^2\bar{x}^2/\sigma_x^2]\right\}}{2 + \frac{1}{n}\left\{1 + [(n^*-1)^2\bar{x}^2/\sigma_x^2]\right\}} \quad . \qquad (2.12)$$

If n is allowed to increase while $(\bar{x}/\sigma_x)^2$ is held fixed, $h_{n+1}/(2 - h_{n+1})$ will remain relatively constant if n^* is proportional to $n^{\frac{1}{2}}$. Thus, in order to maintain a constant relative leverage, the augmented point must maintain a distance of approximately $n^{\frac{1}{2}}$ from the original data.

2.4 Calculational Considerations

By augmenting the data set with the point $(n^*\bar{x}, n^*\bar{y})$, a computer program which does not have the option to fit the no-intercept model can be used to fit such a model. Hawkins (1980), and Criner and McElroy (1976), independently noted that if the original data is augmented with n new data points consisting of the reflections in the origin of the original data, then the full fit on this new data set is equivalent to forcing the regression through the origin. The associated statistics are correct except for some minor modifications, which are explained in both papers.

If the point $(n^*\bar{x}, n^*\bar{y})$ is added to the original data and the full fit is performed using a statistical computer package, the associated statistics (and degrees of freedom) are correct for the no-intercept model. In particular, the outputted R^2, F, and the residual mean square are the correct quantities for the no-intercept model **(although, as noted before, care must be taken when using some of these statistics)**. This method of adapting a program to perform the fit through the origin will, in some cases, be easier to implement than the others, especially if the calculations are being done with a hand-held calculator. Many calculators today have built-in linear regression, but usually without the no-intercept option. It is relatively easy, however, to calculate the augmented point (since Σx and Σy will be in storage) and add it to the original data set.

3. NUMERICAL EXAMPLE

To illustrate the leverage point interpretation of regression through the origin, consider the gas mileage data of Hocking (1976), which is reported in full in Henderson and Velleman (1981). Only two variables will be considered here: gallons per mile (GPM), which is the inverse of miles per gallon, is the dependent variable, and total weight of the vehicle (WT) is the independent variable.

Henderson and Velleman suggest that WT is the best single predictor of GPM, and that the relationship is very close to linear. (A plot of miles per gallon vs. WT shows a distinct nonlinear relationship.) Furthermore, they provide an informal theoretical argument which shows that there are physical grounds for fitting a line through the origin. Briefly, if gasoline consumed is proportional to the work expended in moving the vehicle, this work is also proportional to the weight of the vehicle, hence GPM α WT. (Theoretically, a vehicle with zero weight will consume zero fuel.) Thus, considering the physical constraints imposed by the relationship between GPM and WT, it seems most appropriate to fit a line through the origin. However, for purposes of prediction, it may be desirable to assess the adequacy of both the intercept and no-intercept model.

The data set, taken from 1974 Motor Trend magazine, contains measurement on 32 different automobiles. The variable GPM is rescaled to gallons per 100 miles, a more convenient set of units, and WT measures weight in 1,000 lbs. The full fit to these data yields the least square equation $\text{GPM} = .617 + 1.494\text{WT}$, with $R^2 = .792$. The fit through the origin gives $\text{GPM} = 1.67\text{WT}$, with $R^2 = .982$. At first glance, the higher R^2 of the regression through the origin suggests that this is the better fit. However, it may be the case that this value is artificially inflated, since it is based on the uncorrected total sum of squares.

The alternate method for calculating R^2 for the regression through the origin, suggested by Gordon (1981), is

$$R_G^2 = 1 - (\text{Residual SS}/\text{Corrected Total SS}) \quad , \qquad (3.1)$$

where Residual SS is the residual sum of squares based on the model being fitted (in this case Residual $SS = \Sigma y_i^2 - [(\Sigma x_i y_i)^2/\Sigma x_i^2]$), and the Corrected Total $SS = \Sigma(y_i - \bar{y})^2$. For the no-intercept model, $R_G^2 = .780$, showing that the intercept and no-intercept models fit the data equally as well. Indeed, Figure 1, a scatterplot of the data together with the two least squares lines, shows the models to be virtually indistinguishable.

The hypothesis H_0: $\alpha = 0$ in the full model has $t = .698 (p > .4)$, and hence supports the conclusion that the no-intercept model is appropriate. Thus, taking into account that the physical constraints support the no-intercept model, and that the statistical evidence shows this model to be adequate, the regression through the origin seems to be the better choice.

When the original data are plotted together with the augmented point, as in Figure 2, a slightly different picture is seen. The augmented data point $(n^*\bar{x}, n^*\bar{y}) = (21.697, 36.576)$ is a point of high leverage ($h_{n+1} = .920$), but does not affect the fit in an adverse way ($r^*_{n+1} = .698$) . The regression through the origin, acting as if a full fit is being done to the augmented data set, "sees" two distinct clusters; the original data and $(n^*\bar{x}, n^*\bar{y})$. The least squares line from the no-intercept model virtually goes through the augmented point.

The vertical distance between the augmented point and $\hat{\alpha} + \hat{\beta}(n^*\bar{x})$, the predicted value at $n^*\bar{x}$ from the full fit on the original data, is

$$|n^*\bar{y} - [\hat{\alpha} + \hat{\beta}(n^*\bar{x})]| = (n^* - 1)|\hat{\alpha}| \quad , \qquad (3.2)$$

showing that the discrepancy between the two models, at the augmented point, is proportional to $|\hat{\alpha}|$ (which measures the discrepancy at the origin). This magnification of the discrepancy between the two models can be a useful visual aid, and can help to choose the more appropriate model.

If prediction is a major concern, Figure 2 also provides guidance in deciding between models. It is sometimes possible to assess if the augmented data point could, in fact, be a valid data point. That is, using his expertise and knowledge of the physical situation, an experimenter might be able to decide if it is reasonable for an observation taken at $n^{*}\bar{x}$ to yield a response close to $n^{*}\bar{y}$. Indeed, in some cases, it may be possible to take such data experimentally. In either case, assessing whether $(n^{*}\bar{x}, n^{*}\bar{y})$ might be a valid data point can help in deciding which model is the better predictor.

For the gas mileage data, the augmented point corresponds to a vehicle weight of 21,697 lbs. and 2.73 miles per gallon. This value of weight is beyond those of automobiles, but is close to values for large trucks. If there is interest in extending the predictions to vehicles of such weights, we should try to decide if 2.73 miles per gallon is a reasonable value. (There is, of course, a problem in that most trucks use diesel fuel, while all but one of the automobiles in the original data were gasoline powered. However, we might just as well ignore this, since the major question is which of these simple models perform better realistically.) The value of 2.73 miles per gallon is less than is to be expected from a truck of approximately 21,000 lbs., and hence a value lower than 36.576 GPM is more reasonable. Thus, for predictions at these higher weights, the intercept model should provide more reasonable results than the no-intercept model.

REFERENCES

CRINER, J. C., and McELROY, F. W. (1976), "Estimation in Regression without a Constant Term Using a Computer Program Which Assumes There is One," *Proceedings of the Section on Statistical Computations of the American Statistical Association*, 137.

GORDON, H. A. (1981), "Errors in Computer Packages. Least Squares Regression Through the Origin," *The Statistician*, 30, 23-29. (See also the Letter to the Editor by P.L. Goldsmith, and subsequent discussion by H.A. Gordon in *The Statistician*, 30, 304-308.)

HAHN, G. J. (1977), "Fitting Regression Models with No Intercept Term," *Journal of Quality Technology*, 9, 56-61.

HAWKINS, D. M. (1980), "A Note on Fitting a Regression without an Intercept Term," *The American Statistician*, 34, 233.

HENDERSON, H. V., and VELLEMAN, P. F. (1981), "Building Multiple Regression Models Interactively," *Biometrics*, 37, 391-411.

HOAGLIN, D. C., and WELSCH, R. E. (1978), "The Hat Matrix in Regression and ANOVA," *The American Statistician*, 32, 17-22.

HOCKING, R. R. (1976), "The Analysis and Selection of Variables in Linear Regression," *Biometrics*, 32, 1-44.

MALLOWS, C. L. (1973), "Some Comments on C_p," *Technometrics*, 15, 463-481.

MARQUARDT, D. W., and SNEE, R. D. (1974), "Test Statistics for Mixture Models," *Technometrics*, 16, 533-537.

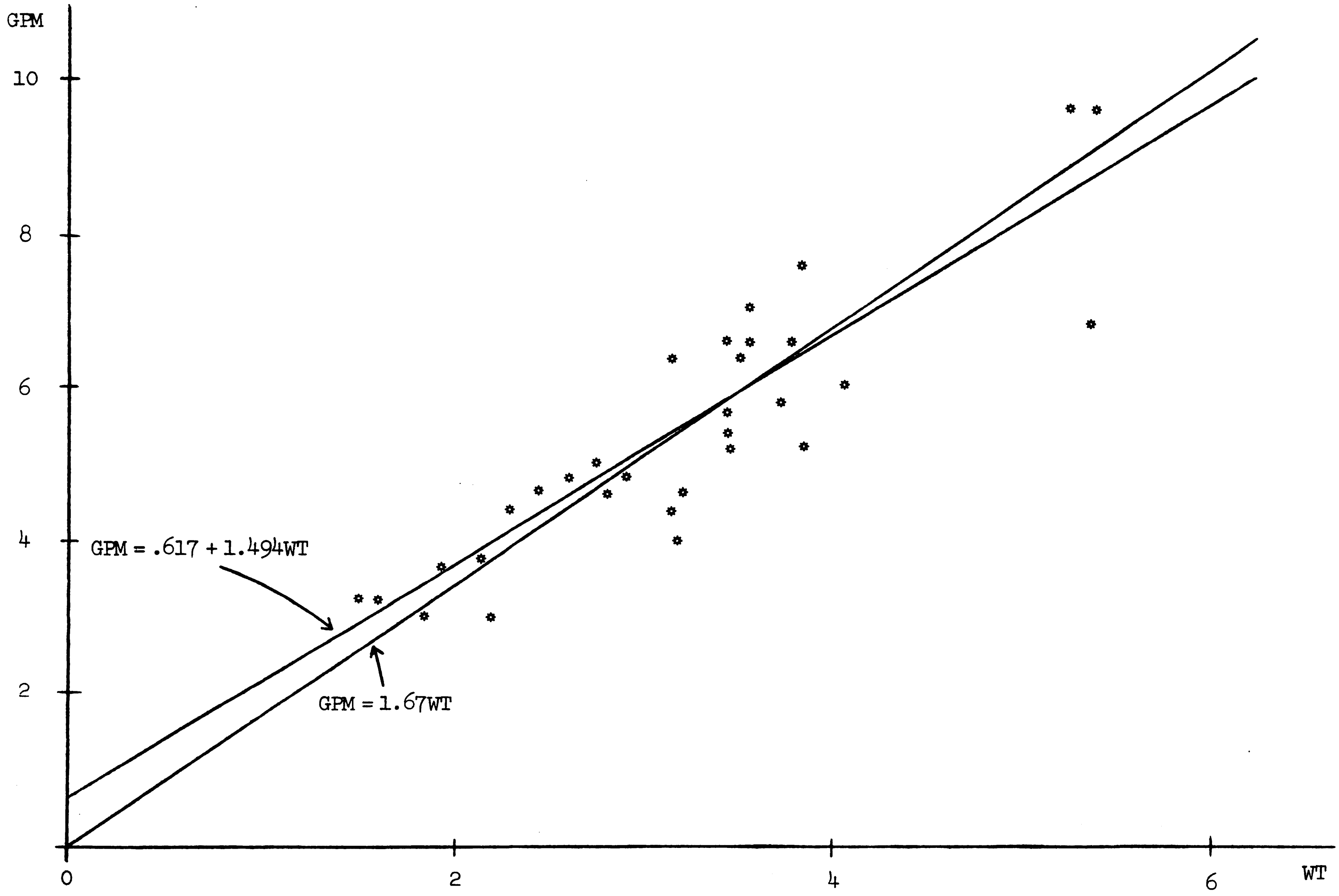

Figure 1: The Original 32 Observations of the Gas Mileage Data.

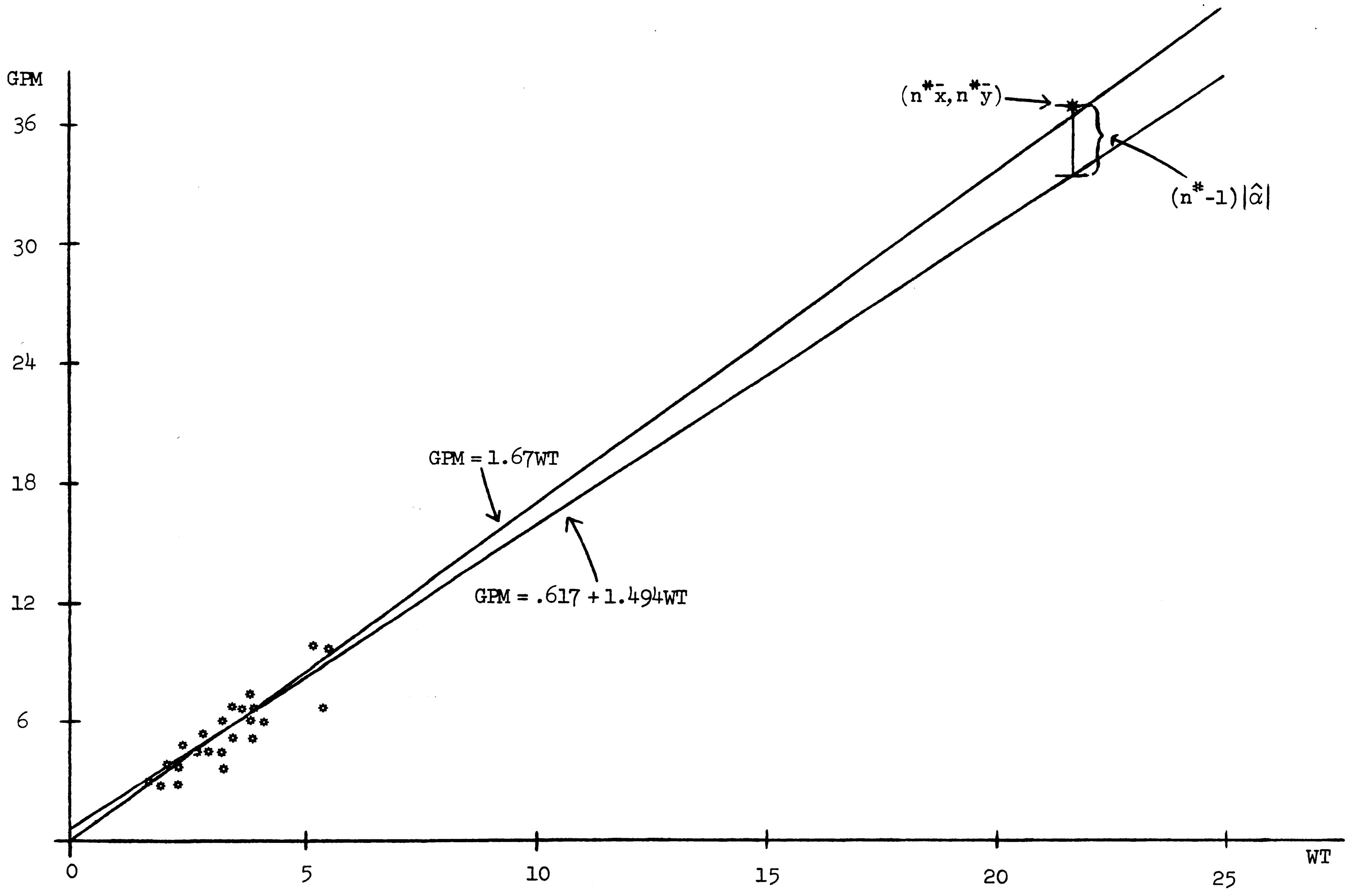

Figure 2: The Original Observations Plus the Augmented Point $(n^*\bar{x}, n^*\bar{y}) = (21.697, 36.576)$.