1981

# The Use of Multiple Regression Residuals Analysis in Restructuring Prediction Equations

Mary E. Malliaris
*Loyola University Chicago*

# THE USE OF MULTIPLE REGRESSION RESIDUALS ANALYSIS IN RESTRUCTURING PREDICTION EQUATIONS

by

Mary E. Malliaris

A Dissertation Submitted to the Faculty of the Graduate School
of Loyola University of Chicago in Partial Fulfillment
of the Requirements for the Degree of
Doctor of Philosophy

May

1981

# ACKNOWLEDGMENTS

I would like to thank Dr. Jack Kavanagh, the Director of my Dissertation Committee, for his support, encouragement and guidance in this research study. I also wish to thank Dr. John Wozniak, Dr. Samuel Mayo and Dr. Ronald Morgan for their constructive advice and experienced critical judgement during the preparation of the manuscript.

I would like to thank Dr. Donald Meyer, Dean, School of Business, Loyola University of Chicago, for permission to use data from the Graduate School of Business for this study. I also want to thank Glen Emerson for the accurate punching and proofreading of the data.

I am grateful for the patience and emotional support given by my husband during the duration of the project, and the prayers of my mother.

Finally, I want to thank Hugh and Mary Elizabeth Emerson for their emphasis on the value of education and the opportunities they have given me to learn throughout my lifetime.

# VITA

The author, Mary E. Malliaris, is the daughter of Glen D. Emerson and Noma (Nichols) Emerson. She was born October 7, 1948, in Ada, Oklahoma.

Her elementary education was obtained in public schools in the state of Oklahoma in Antlers, Midwest City, Stillwater, Hugo and Muskogee. Her secondary education was obtained at Alice Robertson School and Central High School in Muskogee, Oklahoma. In 1966, she was elected a member of the National Honor Society.

In June, 1966, she entered Louisiana State University in Baton Rouge, Louisiana and received a Bachelor of Science degree with a major in mathematics in January of 1970. While attending Louisiana State University, she was elected a member of several honor societies, including Alpha Lambda Delta, Mu Sigma Rho, Pi Mu Epsilon and Phi Kappa Phi.

In September of 1973, she received a fellowship from Northwestern University, Evanston, Illinois. In June of 1974, she was awarded the Master of Science degree in mathematics.

In 1976, she held a graduate assistantship at Loyola University of Chicago. From 1977 to 1979, she taught mathematics at Loyola University of Chicago.

# TABLE OF CONTENTS

## LIST OF TABLES

# LIST OF FIGURES

Figure Page

Figure Page

Figure Page

Figure Page

Figure Page

Figure Page

# CHAPTER I

## THE INTRODUCTION

Multiple regression analysis is frequently used in formulating equations to predict student performance. The resulting multiple correlations, however, are frequently too low to be of much use in accurate predicting. The purpose of this research was to develop optimal prediction equations for student performance in the Graduate School of Business, based on information available prior to admittance. The traditional multiple regression method was augmented by regressions altered by transformations, additional polynomial variable terms and elimination of outliers. These alterations were determined through the analysis of residuals plotted against the fitted values and each independent variable. The equations were then analyzed and combined to develop the best pair of equations, one for females, one for males, to predict student performance.

This study departed from previous graduate performance prediction studies in the following ways: first, the traditional method of prediction uses the independent variables in the regression equation as they occur. The variables are not checked to see if they actually fit the model requirements or if they might be altered in some way to get better results from the regression model. This approach may be obscuring the actual relationships between the dependent and independent variables. In this study, each variable was adjusted, as necessary, through the use of transformations, elimination of outliers and addition of polynomial variable forms. Secondly, separate prediction equations were developed for males and females rather than one regression equation for the entire Graduate School of Business. This was necessary not only due to the current interest in female performance, but also due to evidence which suggests that traditional indicators of scholastic performance may function better for females and that variables may not differentiate identically for males and females. Thirdly, most research has concentrated on testing more variables, such as personality measures and locus of control, many of which are impractical for the average business school to collect. It was our purpose to improve the prediction by gleaning more information from data currently available to graduate schools of business, rather than attempting to change the application process or to subject applicants to further testing. Lastly, the dependent variable was the entire graduate grade point average, rather than the usual first year average or first semester average. This entire average was more representative of a student's performance, especially at Loyola, where most of the students are part-time and may take up to five years to complete the sixteen course program.

The methodology followed was the generation of multiple regression equations and a following analysis of standardized residuals graphs based on the regression equations. The plots were analyzed for possible model violations. Procedures used to improve the fit included the elimination of outliers, the addition of terms and the use of a transformation. The standardized residuals were examined at several stages to insure that existing violations were being corrected for before a final equation was selected as the optimal one for each group.

The sample for this study was limited to students who received degrees from the Graduate School of Business of Loyola University of Chicago during 1979 and 1980. The results of this dissertation are not extendable to other programs where the majority of students are not part-time. The conclusions of this study are limited by the size, the nature and the time dimension of the sample. The sample included two hundred eighty males and one hundred twenty females, approximately 90% of whom were part-time, taking no more than two courses per quarter. Extending the conclusions of this study to other situations is inappropriate. However, this dissertation provides a direction for a methodology which could be used to improve predictive accuracy in other situations.

# CHAPTER II

## THE REVIEW OF THE LITERATURE

The methods of multiple correlation and regression analysis are frequently used in the process of academic prediction (Misanchuk, 1977, Fishman and Pasanella, 1960, Cooley, 1971, Rao, 1973). Predictor studies have indicated that regression equations function reasonably well and can also aid the admissions committee to resolve cases with conflicting applicant information (Boldt, 1969, Dawes, 1971). Multiple variables are preferred since, as Wiggins (1968) points out, "the probability that one dependent variable (or variation thereof) has multiple causes (independent variables) is greater than the probability that it is caused by a single independent variable" (p. 390). When combining several scores, Weinstein, Brown and Wahlstrom (1979) argue

> In the composite score procedure, the scores of several tests are combined by an algorithm to yield a single score. There are three basic types of composite score procedures: unweighted, opinion weighted, and regression weighted. ...
>
> The regression-weighted composite is the most effective approach, insofar as it is the best predictor of student success if a linear relation exists between test scores and the criterion measure. (p. 130)

Engelhart (1972) favors using regression equations when the criterion variable is "to some extent unreliable", as the case often is with prediction of grades (p. 324).

When using regression to predict academic performance, Mosteller and Tukey (1977) state that the prediction clearly needs "to be at least partly successful" (p. 270). But this is often not the case. Though a very popular approach, correlation and linear regression have not yielded high predictive accuracy (Dawes, 1975, p. 721). Givner and Hines (1979), in discussing the correlation of scores on admissions tests with school performance state that experience "reveals that such correlations are often far lower than what was intuitively expected" (p. 119). Pitcher and Smith obtained an average multiple correlation of .43 in predicting the first year average from undergraduate record and the Advanced Test for Graduate Schools of Business (ATGSB) scores for full-time day students and an average multiple correlation of .30 for students from mixed full-time, part-time, day and evening programs. With the addition of interruption as a dummy variable, Pitcher (1973) obtained an average multiple correlation of .43. In a study of fourteen graduate schools of business, Pitcher and Schrader (1972) reported an average multiple correlation of .44 using undergraduate record, ATGSB verbal and ATGSB quantitative scores as independent variables. Burnham and Hewitt ( 1972) got a multiple correlation of .51 between College Board Scores, high school average, and achievement in college. Schwartz and Clark (1959), in a study of graduate success at Rutgers University, found a multiple correlation of .43 between graduate averages and undergraduate grade point averages, the Miller Analogies Test and the Doppelt Mathematical Reasoning Test. This led them to conclude that "none of the correlations is high enough to give strong confidence in any of the predictors" (p. 111). Petry and Craft (1976) obtained a correlation of .49 in predicting grade point average from the Cooperative School and College Ability Test, Verbal, Mathematical and Total scores. These low correlations may be due to the fact that the group available for measuring is restricted. Givner and Hynes (1979) state that "the more restrictive the accepted group on the admissions test, the greater the resulting correlation coefficient will underestimate the actual correlation in the nonrestricted group or population" (p. 120). They may also reflect upon the validity of standardized exams required for admittance to most programs. As Ebel (1972) states, "even the list of available published tests should be regarded with at least two grains of salt: justifiable skepticism about what some of the tests actually measure, and about the quality of published tests" (p. 453). Weinstein adds that "transcripts offer a four year sample of academic abilities, while tests offer a sample of just a few hours" (p. 135).

Another major cause of low correlations may be the use of linear regression without in-depth analysis of the variables to detect and correct model violations and misspecifications. The graphical analysis of the residuals is necessary in order to discover and correct these violations. Tukey (1977) states that "one of the great arts of data analysis consists of subtracting out incomplete descriptions and examining the residuals that are left" (p. 143). A large value of $R^2$ or a significant F statistic does not insure that the model is correct and the data has been fitted well. To emphasize this fact, Anscombe (1973) has given four sets of data, all having identical summary statistics, but each with a different and distinct pattern.

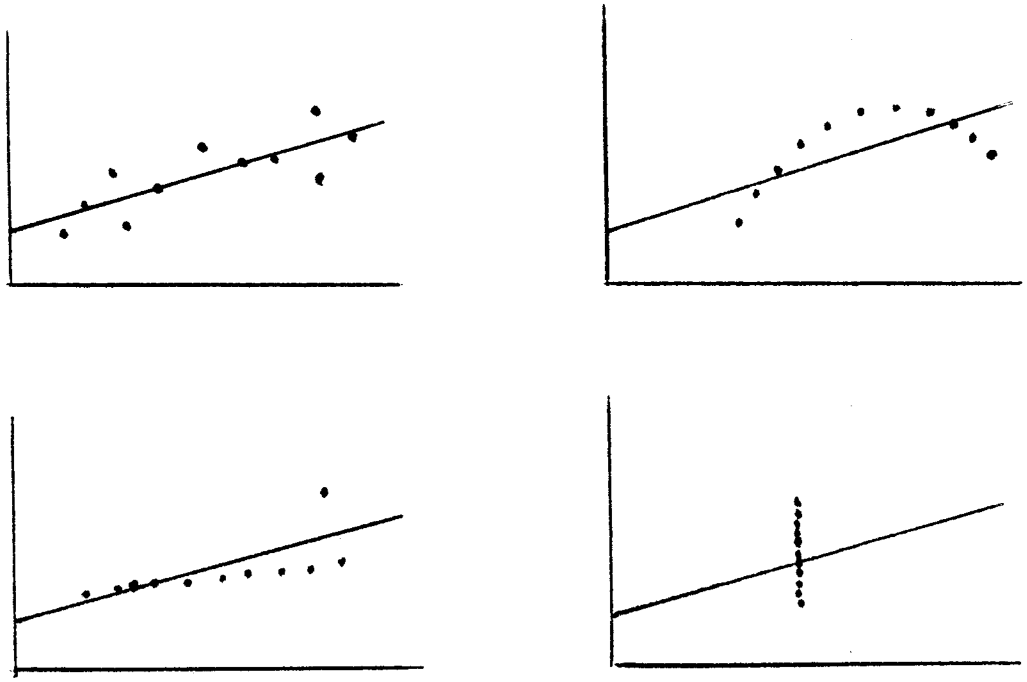
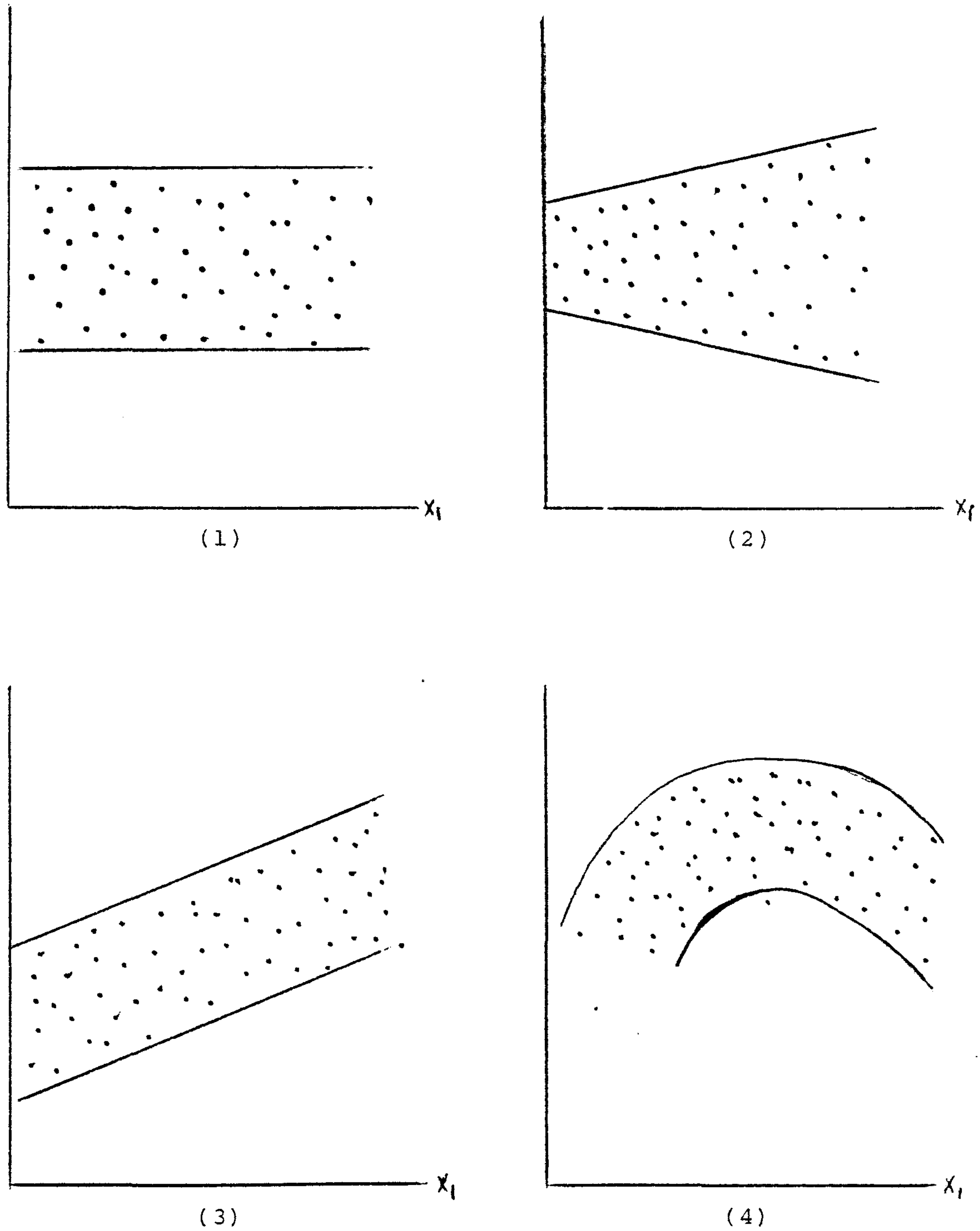


Figure 1. Four data sets with identical summary statistics (Anscombe, 1973, p. 19).

An analysis based solely on summary statistics would not have been able to discern these differences. A poor fit in terms of a low $R^2$ or a non-significant F may occur not because of a large error term, but because the model was fitted to a nonlinear relationship. Residual analysis not only reflects the true error but also the variation of a true curvilinear relationship from an assumed linear one.

Residual analysis is a graphical method which includes the examining of plots of standardized residuals against the fitted value $\hat{y}$ and each of the independent variables $x_i$. Standardized residuals have a mean of zero and a standard deviation of one. Generally, if the model is correct, the standardized residuals fall between plus and minus two and are randomly distributed about zero (Chatterjee and Price, 1977, p. 10). If the model is invalid, the residual plots will show a distinct pattern or may fall outside the specified range. Figure 2, from Draper and Smith (1966, p. 89), indicates several patterns which may occur when the standardized residuals are plotted against any of the variables. The first plot, (1), shows a proper fit. Plot (2) shows a variance which increases as the value of the variable increases, i.e., heteroscedasticity, implying the need for some sort of transformation. Plot (3) shows the need for the addition of a linear term. The last plot, (4), shows a curvilinear relationship which implies the



(1)

(2)

(3)

(4)

Figure 2. Typical patterns of standardized residuals versus $x_1$.

need for quadratic terms in the regression equation. After the residuals plots are examined, we can conclude either that the model assumptions have been violated in a specific way or that the assumptions do not appear to have been violated (Draper and Smith, 1966, p.86). We can use the graphs to check for violations concerning inadequacy of the linearity assumptions, lack of constant variance, presence of outliers, and correlated errors (Chatterjee and Price, 1977, p. 20).

## Outliers

Outliers are extreme data points. "The outlier is a peculiarity and indicates a data point which is not at all typical of the rest of the data" (Draper and Smith, 1966, p. 94). In constructing a regression equation, it is important that the results are not solely dependent on a few observations. Anscombe (1973) states,

> We are usually happier about asserting a regression relation if the relation is still appropriate after a few observations (any ones) have been deleted--that is, we are happier if the regression relation seems to permeate all the observations and does not derive largely from one or two. (p. 18)

Outliers become visible in a scatterplot of residuals. In Figure 3 is a graph of standardized residuals versus the independent variable, from Chatterjee and Price (1977, p. 23). The outliers, shown as circled data points, occur in the upper right-hand and lower left-hand corners of the plot. Elimination of these points reduced the

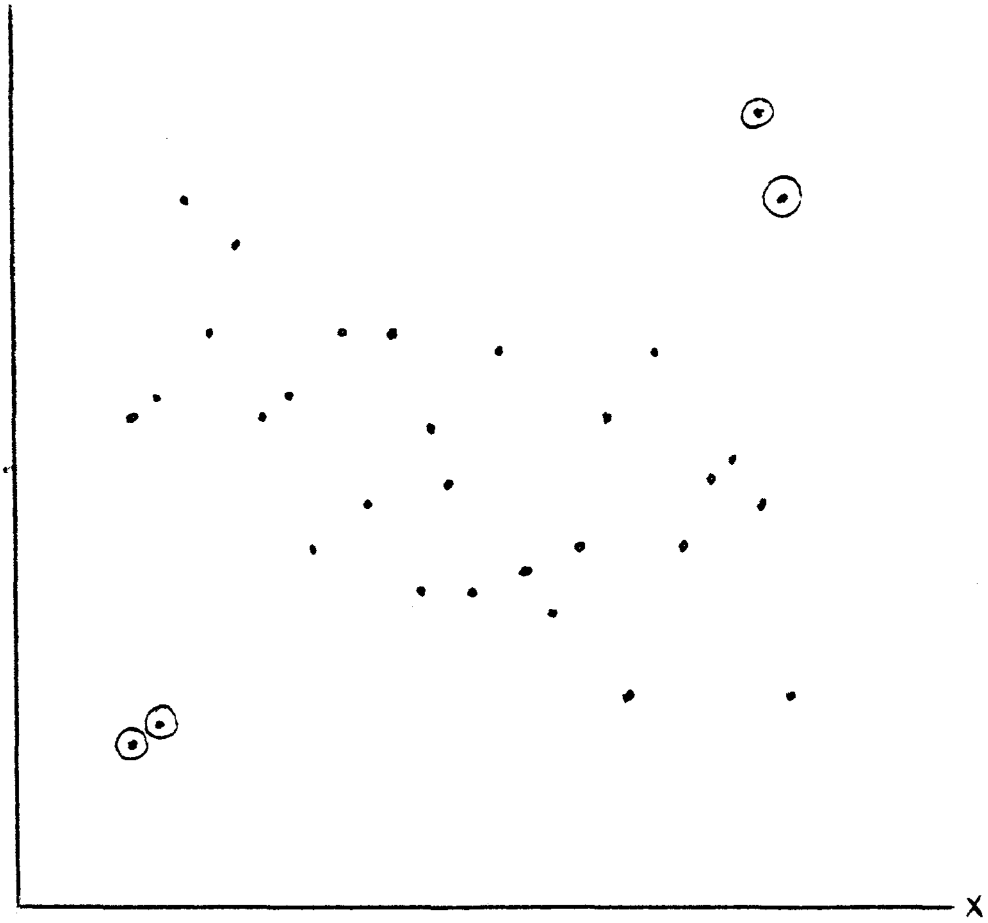


Figure 3. Residual plot showing outliers.

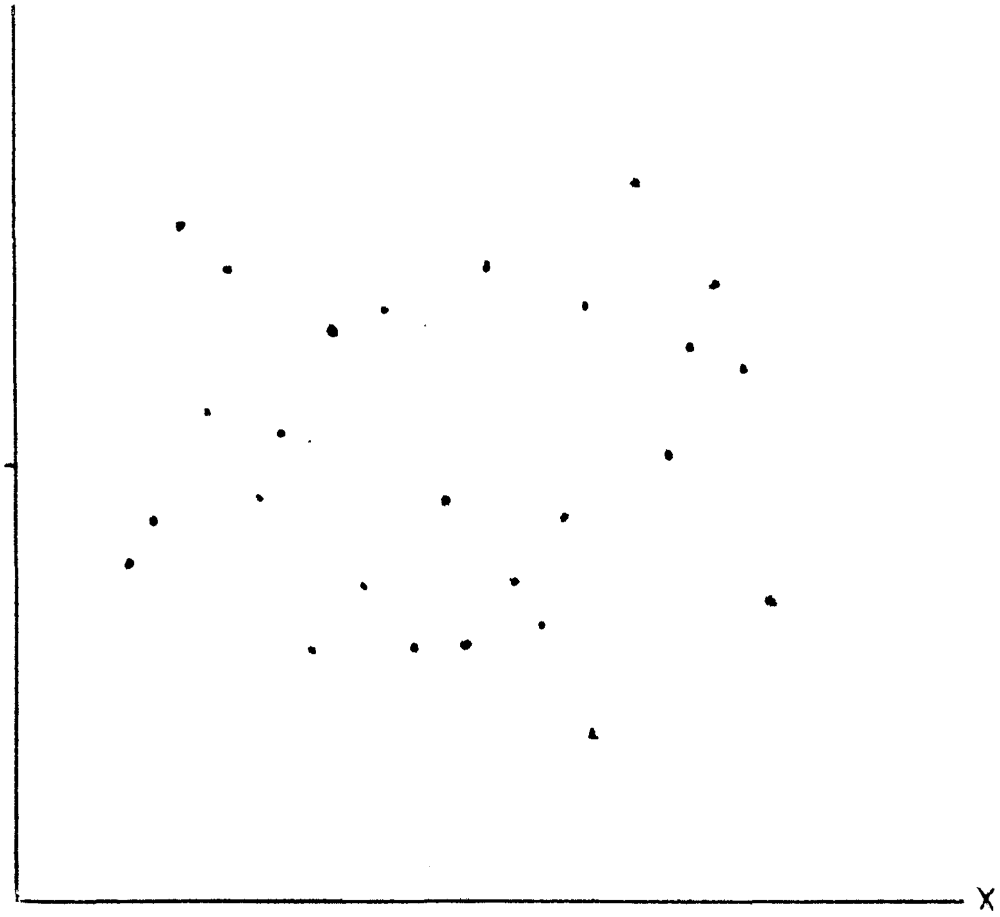


Figure 4. Residual plot with outliers removed.

standard error from .817 to .631. Figure 4 is the plot of standardized residuals versus the independent variable, with the outliers eliminated. The residuals appear to be distributed randomly about zero, which indicates that the regression model is a satisfactory model for the data, after the outliers have been eliminated. Comparison of the regression parameters before and after eliminating the outliers is shown in Table 1.

Table 1. Summary of Regression Results For Full and Reduced Data Sets (Chatterjee and Prive, 1977, p. 26).

| | Full data set | Reduced Data Set |
|---|---|---|
| $b_1$ | 0.665 | 0.260 |
| $b_o$ | 1.706 | 3.713 |
| s | 1.402 | 0.925 |
| $R^2$ | 0.396 | 0.161 |

As the change in the coefficients indicates, the method of estimating parameters is very sensitive to outliers. In the development of an equation to predict typical performance, it is important to exclude grossly atypical points. Otherwise, a very small proportion of the data has an extreme effect on the regression equation and may lead to conclusions quite different from those based on the reduced data set.

Heteroscedasticity

When the error variance is not constant, over all observations, the error is said to be heteroscedastic. Heteroscedasticity is a violation of one of the model assumptions of multiple regression. In Figure 5, we have two plots of standardized residuals versus $x_i$, an independent variable. Plot (a) shows residuals which increase in scatter as the value of $x_i$ increases and plot (b) shows residuals which decrease in scatter as $x_i$ increases. Each of these plots is an indication of heteroscedasticity, or unequal error variance (Daniel and Wood, 1971, p. 28, Chatterjee and Price, 1977, p. 47, Kim and Kohout, 1975, p. 342).

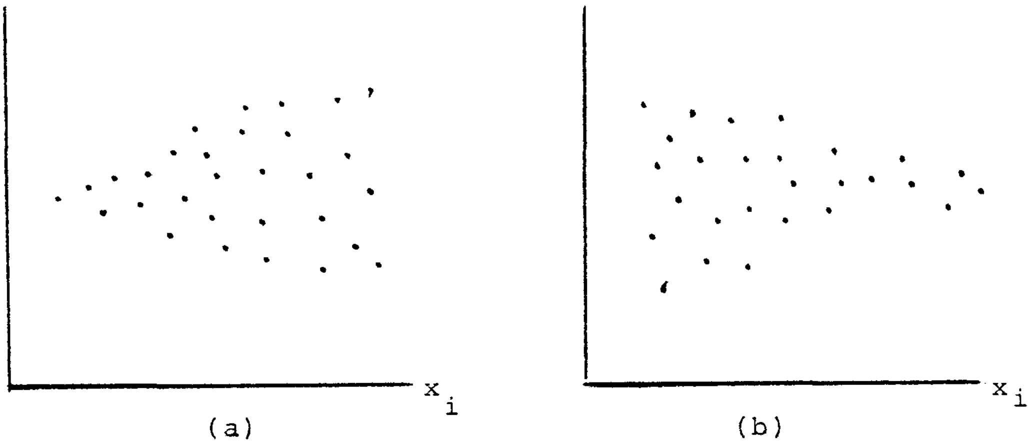


Figure 5. Residual plots indicating heteroscedasticity.

The consequences of heteroscedasticity are the following: the estimators of the regression parameters are still unbiased, however, they no longer have minimum variance. Also, the standard errors are incorrect and as a result, tests of significance and confidence intervals for the regression parameters may be very misleading (Wesolowsky, 1976, p. 126).

Heteroscedasticity can often be removed by using a suitable transformation on the data. In the following example, taken from Wesolowsky (1976, p. 132), the data was first fitted to the equation

$$y = \beta_0 + \beta_1 x_1.$$

A plot of the standardized residuals versus $x_1$ shows a funnel shape with the spread of the residuals increasing as $x_1$ increases (Figure 6).

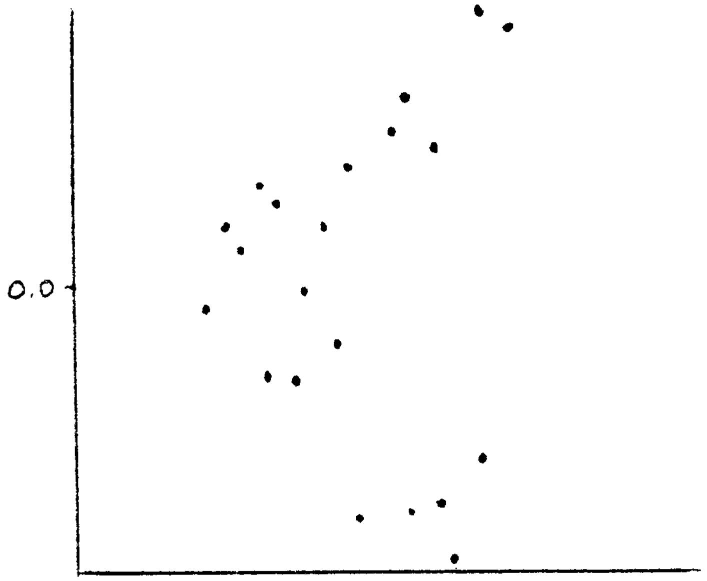


Figure 6. Funnel-shaped plot of residuals.

To restore homoscedasticity, a transformation of $(1/x_1)$ was applied to the equation. the second equation fitted was

$$\frac{y}{x_1} = \frac{\beta_0}{x_1} + \frac{\beta_1 x_1}{x_1}$$

or equivalently,

$$\frac{y}{x_1} = \beta_1 + \frac{\beta_0}{x_1} .$$

The plot of standardized residuals versus $x_1$ in the second run is shown in Figure 7.

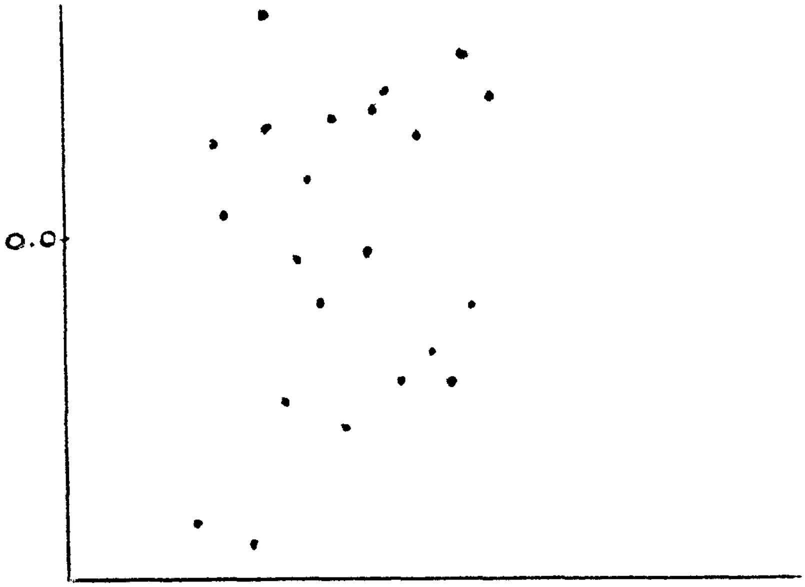


Figure 7. Residual plot after correcting for heteroscedasticity.

The residuals no longer have a pronounced funnel shape, indicating that the heteroscedasticity has been removed.

If heteroscedasticity is such that the variance decreases as $x_i$ increases, Daniel and Wood (1971, p. 27) suggest a transformation where each variable is weighted by

$(1/(V - x_i)^2)$ where V is the value which the variable approaches as the scatter of residuals decreases. This will give different $\beta$ values but a better fit.

## Autocorrelation

Another original assumption of the regression model is that successive error terms are uncorrelated with previous values. Autocorrelation is the statistical dependence of errors on preceding errors. If autocorrelation exists and is not compensated for, the standard errors of the sample regression coefficients underestimate the true standard deviation, i.e., the estimators are biased. $R^2$ may also be higher than it should be and the parameter estimates, though still unbiased, are no longer the minimum variance estimators (Wesolowsky, 1976, p. 137). When errors are correlated, residuals of the same sign tend to occur in clusters or bunches. Several successive residuals are positive, the next several negative, and so on. The graph in Figure 8, from Chatterjee and Price (1977, p. 126), of standardized residuals versus time suggests that the error terms are correlated.

The coefficient of autocorrelation is $\rho$. The test of the hypothesis $H_o$: $\rho = 0$ may be done using the Durbin-Watson statistic, D, where

$$D = (\sum_{i=1}^{n} (e_i - e_{i-1})^2 / \sum_{i=1}^{n} (e_i)^2.$$

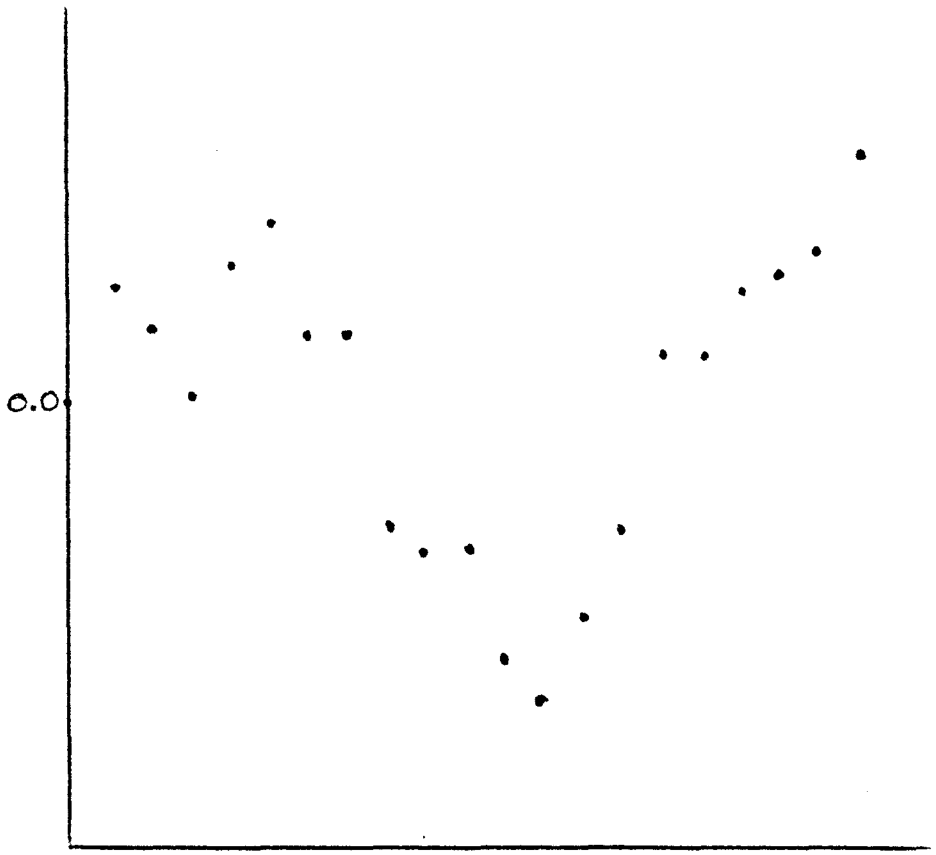


Figure 8. Residual pattern suggesting autocorrelation.

$e_{i-1}$ and $e_i$ are successive errors. D has a range of 0 to 4. The closer the sample value of D is to 2, the firmer the evidence that autocorrelation is not present in the error. The rule for rejecting $H_o$, using the Durbin-Watson tables, is

If $D > d_U$, do not reject $H_o$

If $D < d_L$, reject $H_o$

If $d_L < D < d_U$, the test is inconclusive.

If autocorrelation is found to exist, the first possibility to consider is that the independent variable relationship to y may not be a straight line. A polynomial form or transformation may be necessary to correct for this. For example, the pattern in Figure 9 is likely to give correlated errors. This type of autocorrelation is autocorrelation in appearance only. It is due to omitting a variable, such as $x_i^2$, that should be in the model. Once the variable is added, the autocorrelation problem is resolved. If the difficulty cannot be eliminated this way, an adjusted equation of the form

$$y_i - \rho y_{i-1} = \beta_0(1-\rho) + \beta_1(x_{1i}-\rho x_{1,i-1}) + \ldots + \varepsilon_i - \rho\varepsilon_{i-1}$$

may be tried (Wesolowsky, 1976, p. 144). This equation is fitted using $(y_i - \rho y_{i-1})$ and $(x_{ji} - \rho x_{j,i-1})$ as the variables.

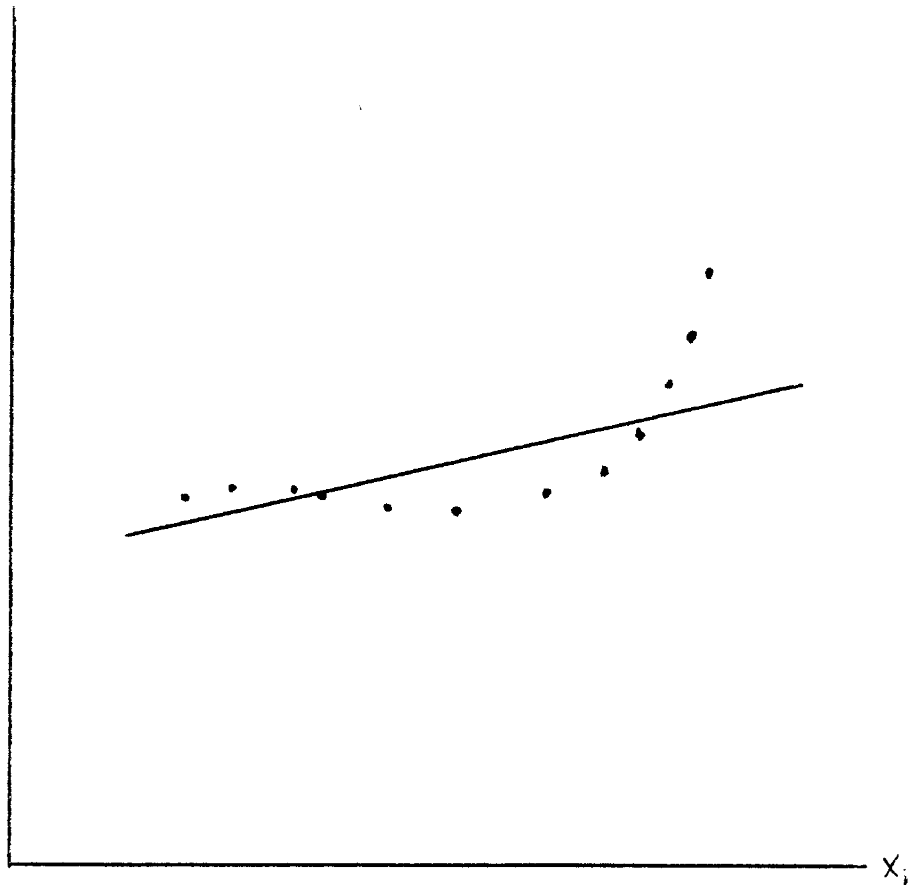


Figure 9. A straight line fitted to an apparently nonlinear relationship (Wesolowsky, 1976, p. 137).

Transformation and Additional Variables

A frequent necessary adjustment in the regression model may be the addition of polynomial terms. A regression model is linear when the parameters occur linearly in the model (Chatterjee and Price, 1977, p. 27). The variables, however, may be transformed in some ways without violating this linearity requirement. For example,

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2^{\;2} + \beta_3 \log x_3 + e$$

is a regression model with the addition of the $x_2^{\;2}$ term and a transformed independent variable. Residual patterns which indicate curvature make the addition or transformation of terms necessary in order to get a proper fit. If the model is not specified correctly, and if we suppose an $x_i^{\;2}$ term is needed, then the plot of standardized residuals against $x_i$ would show a systematic curvature as opposed to a random scatter. Addition of polynomial terms also allows the use of interaction terms when these are of interest to the researcher. Interactive relationships can be investigated by the use of multiplicative terms in the regression equation (Kim and Kohout, 1975, p. 372). If, for example, the researcher believes variables $x_1$ and $x_3$ may be producing an interaction effect, the variable $x_1x_3$ may be added to the equation and treated as an additional independent variable. The equation

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_3$$

is still linear and additive since the $\beta$'s enter the equation linearly. By adding multiplicative terms to the equation, $R^2$ is always increased, though not necessarily significantly (Kim and Kohout, 1975, p. 373). It is necessary to test the null hypothesis that the additional term coefficient is not significant.

Additional terms may be needed when the residuals plot shows that larger observed values are underpredicted by the model (Draper and Smith, 1966, p. 122). Figure 10 illustrates this case. Six out of the seven largest residuals are positive. The addition of variables is necessary to provide better prediction at higher levels.

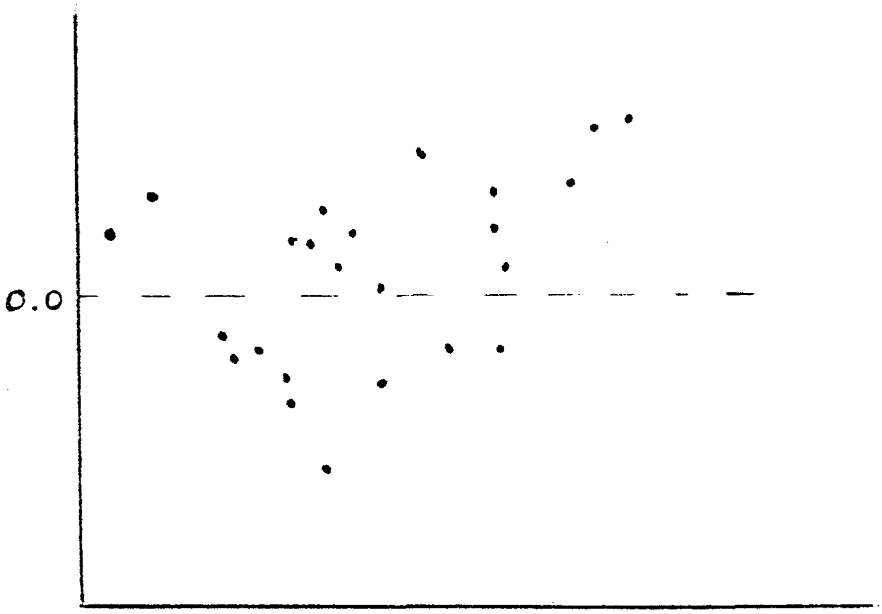


Figure 10. Residuals indicating underpredicted values.

Frequently, qualitative variables can be useful additions to a regression equation. These are known as dummy variables or indicator variables. A dummy variable takes on only two values, zero and one. These values do not reflect a quantitative difference in the variables, but are used only to identify category membership. More than one dummy variable may be inserted into an equation along with several continuous factors (Daniel and Wood, 1971, p. 56). The use of a dummy variable and its effect on the regression equation can be seen in the following example from Wesolowsky (1976, p. 105).

An economist, in a study of fifteen years of the growth of calgacite production, believed a specific year, the ninth, marked an increase in the market for calgacite but that the rate of growth for the market remained unchanged. He introduced this factor as a dummy variable.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \qquad \text{where } x_2 = \begin{cases} 1 \text{ if } x_1 \geq 9 \\ 0 \text{ if } x_1 < 9 \end{cases}$$

and $x_1$ represents the year, y is demand for calgacite in thousands of tons. The resulting equation was

$$y = 9.59 + .757x_1 - 2.11x_2$$

or alternatively,

$$y = 9.59 + .757x_1 \qquad \text{if } x_1 < 9$$

$$y = 7.49 + .757x_1 \qquad \text{if } x_1 \geq 9.$$

Addition of the dummy variable resulted in a shift of slope of 2.10 units and gave a much more accurate prediction equation.

After fitting the equation with a polynomial or a dummy variable, we must test the hypothesis $H_o$: $\beta_i = 0$ and analyze the residual plots. The hypothesis test will tell us if the variable gave us a significant change and the residual plot will indicate whether or not the variable was misspecified.

When a dummy variable of experience was added to a regression equation, Chatterjee and Price (1977, p. 79) show the residual plot in Figure 11. This plot suggests that there may be three or more levels of residuals. The plot is not a simple random scatter. The graph suggests that experience was not treated satisfactorily in the model. It was necessary to add interaction terms and additional dummy variables in order to achieve a symmetrically distributed random scatter plot of residuals.

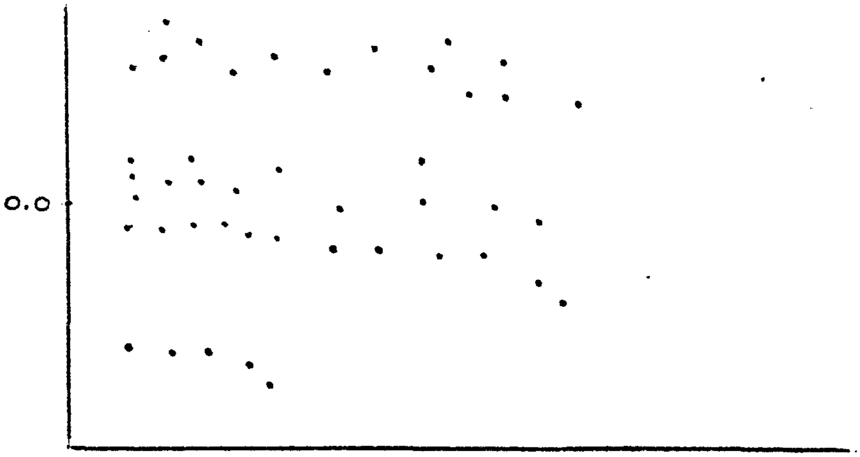


Figure 11. Three levels of residuals.

## Admissions Policies

Admissions policies have come under much scrutiny in the past few years. The federal government, in enforcing equal opportunity, has become involved in what was once considered to be the sphere of individual colleges and universities. The Carnegie Council on Policy Studies in Higher Education (1979) mentions some of the various court decisions which have deeply affected institutional rights:

> In the case of DeFunis v. Odegaard and Bakke v. The Regents of the University of California, the courts have ruled on the appropriate criteria for college admissions. In the case of Goldberg v. Chicago Medical School, the courts ruled on the discretion of colleges to depart from their published admissions criteria. In the case of Barnes v. Converse College, the court ruled on the extent of support services that students must be offered to compensate for deficiencies existing at the time of admission. (p. 50)

Admissions have also been affected by the Education Acts Amendments of the 1964 Civil Rights Act which prohibits "discrimination in college admissions on the basis of a student's sex, race, color, or place of national origin" (The Carnegie Council, 1979, p. 51). Bias may be implicit in the use of specific tests. For example, Maxwell and Jones (1976) state that implicit bias would exist in a "committee decision based solely on GRE-Q scores, where the committee recognizes that male scores tend to be higher than female scores" (p. 33). Discrimination may also result from treating all applicants as one group rather than analyzing them separately (Anastasi, 1976, p. 196). Accrediting agencies have also been active in specifying appropriate policies of individual colleges and refusing to accredit those which do not follow certain standards of admission.

Another factor in concern over admissions policies has been a practical one--the lack of available space. The number of applications is frequently much greater than the number of students which may be admitted. With limited space a definite consideration, schools want to select the best possible set of students to fill the available space. As Weinstein et al (1979) state,

> The place occupied by a student unable or unwilling to complete the program would necessarily be denied another applicant, possibly one more likely to succeed. Therefore, it was imperative to determine whether selection procedures could be identified that gave evidence of ability to filter out potential failures and identify potential successes. (p. 125)

The increasing number of applications also demands an efficient method of analyzing those applications. Dawes (1971) points out that using a paramorphic representation such as a regression equation to select and reject graduate students could save roughly $18 million per year, freeing the professionals on the admissions committees to do more valuable things with their time.

## Variables

As interest in the admissions process and results grows, the variables used in making admissions decisions have been subjected to much analysis, and the search for appropriate variables atill continues. Traditional choices for variables have been the intellective factors of previous scholastic record and scores on one or more standard exams. The effectiveness of these variables as predictors is determined by analyzing their relationship to some measure of success in college, the most common being the student's grade point average (Astin, 1971, p. 3). Fishman and Pasanella (1960) report the "academic grades, and to a less extent, achievement-test scores are strongly entrenched as the criteria of selection and guided admission in American higher education" (p. 306). Non-intellective factors are popular among some researchers. Petry and Craft (1976) found that "studies are about evenly divided between those that investigated intellective variables and those that were concerned with non-intellective variables" (p. 21). Schwartz and Clark (1959), in predicting graduate success, used three predictors, undergraduate grade point average, the Miller Analogies Test, and the Doppelt Mathematical Reasoning Test. Beckham (1973) used the predental grade point average, science courses average and scores on the Dental Aptitude Test to predict performance in dental school. In studies predicting first year averages in graduate schools of business, Boldt (1969) chose undergraduate record and verbal and quantitative scores on the Advanced Test for Graduate Study in Business (ATGSB);

Pitcher and Smith (1969) used undergraduate record and the ATGSB for various subgroups of candidates in twentysix business schools, using age, major field, full-time or part-time status, and a college quality indicator to classify the students. Powers and Evans (1978) used the three standard preadmission variables of undergraduate grade point average, Graduate Management Admissions Test (GMAT) verbal and GMAT quantitative scores to predict the first year average. They found that the GMAT was a more valid predictor than undergraduate grade point average. Pitcher (1973) added a dummy variable of interruption of college studies and a college quality index to the grade point average and achievement test scores. She found that "there seemed to be a general tendency for interrupted students to earn average grades in graduate business school that were higher relative to their measured ability than those earned by uninterrupted students" (p. 3). The most effective single predictors in her study were the ATGSB quantitative and total scores. Pitcher and Schrader (1972) used a college quality index in addition to the standard choices and found that "a small gain in validity can be obtained by using one of the college quality indicators together with the basic combination of undergraduate record, ATGSB Verbal score and ATGSB Quantitative score" (p. 9).

The use of intellective factors has not satisfied all researchers due to the low correlations usually obtained. Many authors have turned from attempting to use traditionally available data to using data based on nonintellective factors. In his lengthy analysis and review of prediction of academic performance, Lavin (1965) points out that

> The relationship between ability and academic performance is well documented and the great majority of studies are no longer concerned primarily with demonstrating this finding. Rather, they attempt to improve predictions through the use of additional factors of a nonintellective nature. (p. 22)

He finally concludes that attention should focus "upon three variables: need for affiliation, need for achievement, and peer group value systems. The first two are personality variables, the third is a sociological characteristic" (p. 162). Casserly and Campbell (1973) suggest including tests of writing ability, oral ability, tolerance for ambiguity, reasoning ability and motivation. Connelly and Nord (1972) propose that "such characteristics as academic exposure, general study habits, attitudes towards business, and personality and motivational characteristics could be studied" (p. 18). Misanchuk (1977) included daydreaming, fear of failure, locus of control, academic values, incentives to achieve, level of work performance, amount of time spent at work, and motivation for alternatives as variables. Gadzella, Cochran, Parham and Fournet (1976) based prediction of grades on students' self-prediction and found that, in some instances, self-prediction is a "better predictor of their academic performance than scores obtained from a mental ability (CTMM) or reading (Cooperative Reading) test" (p. 80). Wikoff and Kafka (1978) also express dissatisfaction with traditional measures of academic potential. They feel that

> Measures of academic potential by themselves are not adequate predictors of academic success. Further studies may be needed to find those personality variables which are most helpful for counseling, but the evidence does seem to indicate that academic success is very much dependent upon personality factors. (p. 323)

Fishman and Pasanella (1960), however, point out that, with regard to using non-intellective criteria, studies must be largely exploratory "inasmuch as no college selects students solely on the basis of motivational and additudinal characteristics of applicants" (p. 303). And Astin (1971), in a study of 135 independent variables to predict freshman grade point averages, found that high school grades were a far more important predictor of grades than aptitude scores and, further, that prediction of academic achievement can be only very slightly improved "by the addition of a wide range of information about the student's family background, race, religion, future plans, and personal attitudes and values" (p. 20).

## Male and Female Performance

Whether one chooses intellective or non-intellective variables, or a combination of them, to predict academic performance, there is evidence to believe that many variables do not predict equivalently for males and females. Women perform better in academic settings than males. Astin (1971) has found that

> Women get higher grades than men both in high school and in college. The academic performance of the female freshman surpasses that of the average male freshman, even when they are matched in terms of their high school grades and aptitude test scores. (p. 20)

In the Maxwell and Jones (1976) study of four graduate programs at the University of North Carolina, it was shown that

> Without exception, the mean grade point average for women applicants is higher than that for males for each program and each year. On GRE scores, women applicants show higher mean scores than male applicants on the verbal test (except for applicants to the Department of English), while males show higher means than females on the quantitative test. (p. 29)

Wikoff and Kafka (1978) have reported that "different variables were found to be discriminating for men and women" (p. 323).

In 1978, there were 40 million working women (Mitchell, 1979) and as the number of women entering the work force increases, the number entering graduate business schools also increases. Not analyzing their performance separately may seriously bias any equation purporting to predict academic potential. Anastasi (1976, p. 192) points out that regression weights may vary from subgroup to subgroup and should be checked whenever there is reason to suspect that a difference may exist. Lavin (1965) found that female performance is "more nearly in accord with their measured ability than is the case for males" (p. 128). In support of treating sex as an important factor, he states

> Ability and school performance are more highly correlated for females than for males. In addition, the absolute level of performance tends to be higher for females. This means that when males and females are not separated in analysis, the magnitude of correlations between ability and school performance will not accurately reflect the true level for the sexes separately.
>
> In the second place, the variables that predict performance for males may be different from the variables that are predictive for females, and even if the same variables are involved for both sexes, the direction of the relationships might differ. (p. 44)

## Summary

Admissions to college have come under much scrutiny. Governmental regulations and increasing enrollments have necessitated analysis of the admissions process and the variables used in decision making. Multiple regression is a frequently used model in forming equations to predict student performance. However, without an accompanying analysis of the residuals, we cannot know if the model is correct. Graphical residual analysis not only spots model violations but also can be a guide to improving the prediction equations through elimination of outliers, correction for heteroscedasticity and autocorrelation, transformation of variables and addition of variables. The most frequently used variables in predicting academic success are previous scholastic record and scores on standardized exams. In admissions to graduate schools of business, these are the undergraduate grade point average and the ATGSB (GMAT). Other variables, intellective and non-intellective are often combined with these. Studies, however, have failed to develop separate equations for males and females.

# CHAPTER III

# THE MODEL

Multiple regression is a statistical technique with which one can analyze the relationship between a dependent variable and a set of independent variables. It is a compensatory model (Wilson, 1973) in that entities with a lot of $x_1$ but a little $x_2$ can have the same y value as entities with a little $x_1$ and a lot of $x_2$ or with moderate amounts of both $x_1$ and $x_2$. Regression is especially useful in bringing out relations between variables whose relation is imperfect in the sense that we may have more than one y for each x (Mosteller and Tukey, 1977).

The data consists of n observations on the dependent variable,y, and each of the p independent variables, $x_1$, $x_2$, ..., $x_p$. The relationship is formulated as a linear model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon$$

where y is the dependent variable, the $x_i$ are independent variables, the $\beta_i$ are constants called regression coefficients or regression parameters and $\varepsilon$ is the error term (or random disturbance). The $\beta_i$ are common to all n observations. $\varepsilon$ includes both measurement error in y and errors in selection of predictors. Any individual value,

$y_i$, is explained by

$$y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip} + \varepsilon_i$$

where $\varepsilon_i$ measures the discrepancy in the approximation for the ith observation. The regression coefficients may be interpreted as the increment in y corresponding to a unit increase in $x_i$ when all the other independent variables are held constant.

The adjective linear implies, in its most general sense, that the regression coefficients enter the equation in a linear fashion. Thus, relationships of the form

$$y = \beta_0 + \beta_1 x_1^2 + \varepsilon$$

and

$$y = \beta_0 + \beta_1 \log(x) + \varepsilon$$

are linear relationships while

$$y = \beta_0 + e^{\beta_1 x} + \varepsilon$$

is a nonlinear model because $\beta_1$ does not enter the model in a linear form but rather in an exponential form (Chatterjee and Price, 1977). A regression model is linear if all $\beta_i$ are raised only to the first power and are not transformed (Finn, 1974, p. 92). It is assumed that the $\varepsilon$'s are random quantities, independently distributed with mean zero and constant variance $\sigma^2$. The $\varepsilon$'s are generally assumed to be mutually uncorrelated and normally distributed for purposes of formal statistical inference.

The estimates of the $\beta_i$ are designated as $b_i$ and are found by minimizing the sum of squared residuals (the method of least squares). Finding the $b_i$ is equivalent to minimizing

$$\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 = \sum_{i=1}^{n}\left[y_i-(\beta_o+\beta_1 X_{1i}+\ldots+\beta_p X_{ip})\right]^2$$

The least squares estimated $b_o$, $b_i$, ..., $b_p$ which minimize are given by the solution of the system of normal equations:

$$s_{11}b_1 + s_{12}b_2 + \ldots + s_{1p}b_p = s_{y1}$$

$$s_{21}b_1 + s_{22}b_2 + \ldots + s_{2p}b_p = s_{y2}$$

$$\vdots$$

$$s_{p1}b_1 + s_{p2}b_2 + \ldots + s_{pp}b_p = s_{yp}$$

where

$$s_{ij} = \sum_{K=i}^{n}(x_{ik}-\bar{x}_i)(x_{jk}-\bar{x}_j) \qquad i,j = 1,\ldots,p$$

$$s_{yi} = \sum_{K=i}^{n}(y_k-\bar{y})(x_{ik}-\bar{x}_i) \qquad i = 1,\ldots,p$$

$$\bar{x}_i = \frac{\sum_{K=1}^{n} X_{ik}}{n} \qquad \bar{y} = \frac{\sum_{K=1}^{n} y_K}{n}$$

and

$$b_o = \bar{y} - b_1\bar{x}_1 - b_2\bar{x}_2 - \ldots - b_p\bar{x}_p$$

Using the estimated regression coefficients, we define the predicted or fitted value

$$\hat{y}_i = b_o + b_1x_{i1} + b_2x_{i2} + \ldots + b_px_{ip}$$

and the observed residual

$$e_i = y_i - \hat{y}_i$$

for each observation. A residual, $e_i$, is the difference between an observed value of the dependent variable and the value predicted by the estimated linear relationship.

Derivations of these equations may be found in Draper and Smith, 1966, Chatterjee and Price, 1977, and Tukey, 1977.

The model is said to fit the data if the variation of the $e_i$ in the sample is small relative to variation in y. Correlation measures reflect the extent to which variation in y is attributable to the independent variables, $x_i$. Altering the order of the independent variables will not affect the estimates of the regression coefficients, although addition or deletion of variables will affect all of the remaining estimates. This is because each estimate is a function of all the $x_i$'s. The set of regression coefficients is determined in order to maximize the prediction of a particular model only.

In matrix notation, the relationships can be expressed in the following way. Let $\underline{y}$ be the (n x 1) vector of dependent variable observations

$$\underline{y} = \begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{bmatrix}$$

$\underline{x}$ is the (n x (p+1)) matrix consisting of a vector of 1's corresponding to the constant term $\beta_c$ and the n values on the p independent predictor variables. $\underline{\beta}$ is the ((p+1) x 1) vector of regression coefficients, and $\underline{\epsilon}$ is the (n x 1) vector of errors.

$$\underline{x} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \cdot & & & \\ \cdot & & & \\ \cdot & & & \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix} \qquad \underline{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \cdot \\ \cdot \\ \cdot \\ \beta_p \end{bmatrix} \qquad \underline{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \cdot \\ \cdot \\ \varepsilon_n \end{bmatrix}$$

The normal equations are generated by

$$(\underline{x'x})\underline{b} = \underline{x'y}$$

and their solutions are

$$\underline{b} = (\underline{x'x})^{-1}\underline{x'y}$$

The assumptions about $\underline{\varepsilon}$ may be written as

$$\underline{\varepsilon} \sim N(\underline{0}, \sigma^2\underline{I})$$

where the expected value of $\underline{\varepsilon}$ is a null vector and $\sigma^2$I is the (n x n) diagonal matrix with $\sigma^2$ in the diagonal and zeros elsewhere. Also, since the expected value of $\underline{\varepsilon}$ is $\underline{0}$, we have

$$\begin{aligned} \mathcal{E}(\underline{y}) &= \mathcal{E}(\underline{x\beta} + \underline{\varepsilon}) \\ &= \mathcal{E}(\underline{x\beta}) + \mathcal{E}(\underline{\varepsilon}) \\ &= \mathcal{E}(\underline{x\beta}) + 0 \\ &= \underline{x\beta} \end{aligned}$$

The variance of $\underline{y}$ for the set of values in $\underline{x}$ is

$$\begin{aligned} \mathcal{V}(\underline{y}) &= \mathcal{E}(\underline{y} - \mathcal{E}(\underline{y}))(\underline{y} - \mathcal{E}(\underline{y}))' \\ &= \mathcal{E}(\underline{y} - \underline{x\beta})(\underline{y} - \underline{x\beta})' \\ &= \mathcal{E}(\underline{\varepsilon}\,\underline{\varepsilon}') \\ &= \mathcal{V}(\underline{\varepsilon}) \\ &= \sigma^2\underline{I} \end{aligned}$$

Let $\underline{b}$ be the ((p+1) x 1) vector of estimates of $\underline{\beta}$, then the properties of $\underline{b}$ include the following: $\underline{b}$ is an unbiased estimator with expected value equal to $\underline{\beta}$; $\underline{b}$ is an efficient estimator, each element being the minimum variance estimate; and if the errors are independent and $\underline{\varepsilon} \sim N(\underline{0}, \sigma^2\underline{I})$, then $\underline{b}$ is the maximum likelihood estimate of $\underline{\beta}$.

It is possible for a regression parameter, $b_i$, to receive a negative value. The multiple regression method selects weights, positive or negative, which give the highest multiple correlation. In the case of negative weights, the variable functions as a suppressor variable to improve the prediction equation. For example, if a test of reading speed is used in conjunction with a speeded history achievement test to predict some external criterion of knowledge of history, the history test is contaminated by reading speed. A negative weight on the reading speed test would help correct for the disadvantage of a student with low reading speed (Darlington, 1968, p. 163).

An unbiased estimate of $\sigma^2$ is

$$s^2 = \frac{\underline{e'e}}{(n-p-1)}$$

where $\underline{e'e} = \sum e_i^2$ is the error sum of squares. $\underline{e'e}$ has (n-p-1) residual degrees of freedom. The estimated standard error of $b_i$ is $\sigma\sqrt{g_{jj}}$, where $g_{jj}$ is the jjth diagonal element of $\sigma^2(\underline{x'x})^{-1}$.

The predicted relationship can be written as

$$\underline{y} = \underline{xb}$$

and the vector of residuals is

$$\underline{e} = \underline{y} - \hat{\underline{y}}$$

If the variation in y is explained to a great extent by the model, we would be more confident of the model's appropriateness. The variation of the dependent variable from its mean can be broken up as follows:

$$(1) \quad \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

(for a derivation, see Wesolowsky, 1976, p. 252) where $\sum(y_i - \bar{y})^2$ is a measure of the variability of the dependent variable; $\sum(\hat{y}_i - \bar{y})^2$ gives the model's contribution to explaining the variation of y about its mean, i.e., the linear relationship's explanation of deviations of y from $\bar{y}$; and $\sum(y_i - \hat{y}_i)^2$ gives the variation unexplained by $\hat{y}$, the fitted linear relationship. The relationships between $\bar{y}$, $y_i$ and $\hat{y}_i$ are illustrated in Figure 12.

Equation (1), restated, is

SS about the mean = SS due to regression<br>
+ SS of residuals

or,

total SS = explained SS + unexplained SS

where the total SS has (n - 1) degrees of freedom, explained SS has (m - 1) degrees of freedom and the unexplained SS has (n - m) degrees of freedom, where m is the number of variables plus 1, m = (p+1).

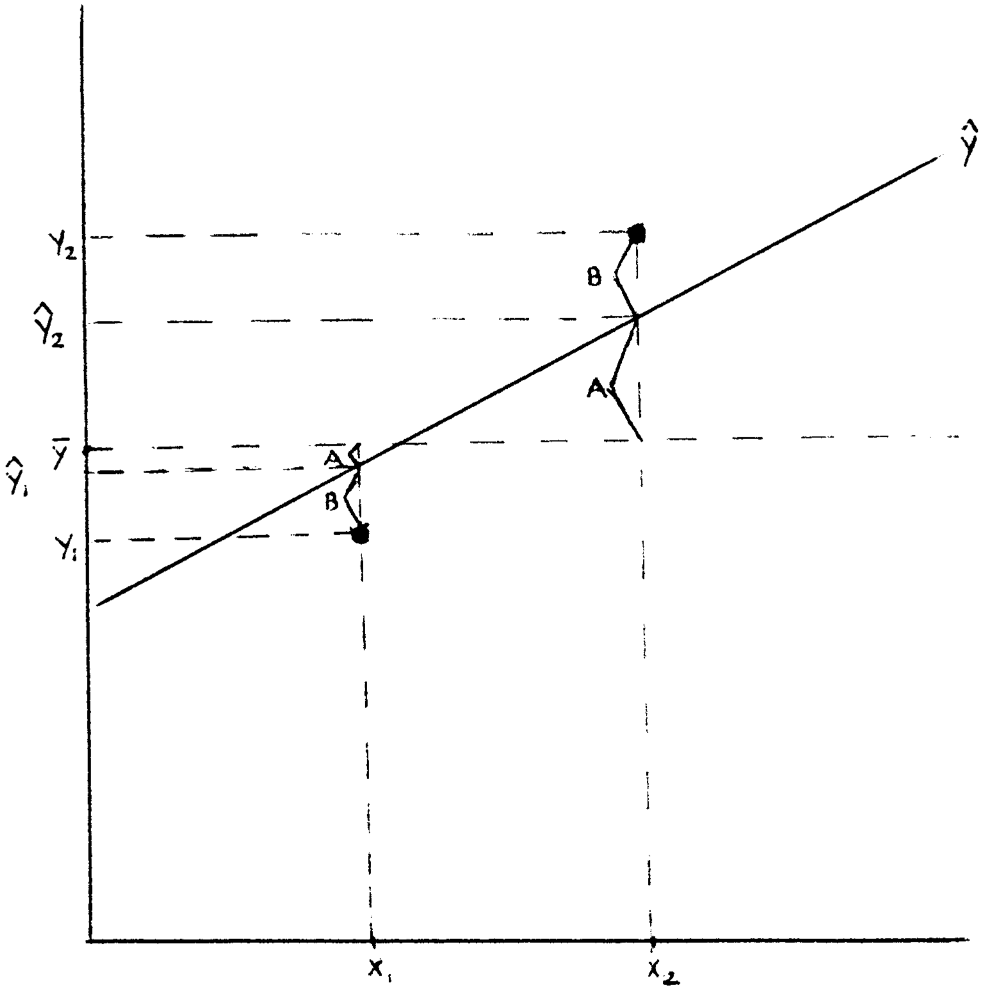


Figure 12. Relationships between $\bar{y}$, $y_i$ and $\hat{y}_i$.

A = variation "explained" by $\hat{y}$

B = variation "unexplained" by $\hat{y}$

Weighted least squares is a method used when the variances of the observations are not all equal. The matrix of observations $\underline{y}$ is transformed to a matrix $\underline{z}$ which satisfies the model assumptions. Let $\underline{y} = \underline{x\beta} + \underline{\varepsilon}$ be the regular model and $\underline{p}$ be a nonsingular symmetric matrix such that $\underline{p'p} = \underline{v}$. If we premultiply the regular model by $\underline{p^{-1}}$ we get a new model

$$\underline{p^{-1}y} = \underline{p^{-1}x\beta} + \underline{p^{-1}\varepsilon}$$

or,

$$\underline{z} = \underline{q\beta} + \underline{f}$$

(see Draper and Smith, 1966, p. 78 for a derivation), and we can apply the basic least squares methods to $\underline{z}$ since $\mathcal{E}(\underline{f}) = \underline{0}$ and $\underline{v}(\underline{f}) = \underline{I}\sigma^2$.

The residual sum of squares is

$$\underline{f'f} = \underline{\varepsilon' v^{-1}\varepsilon} = (\underline{y} - \underline{x\beta})'\underline{v^{-1}}(\underline{y} - \underline{x\beta}).$$

The normal equations are

$$\underline{q'qb} = \underline{q'z}$$

and the solutions for $\underline{b}$ are given by

$$\underline{b} = (\underline{x'v^{-1}x})^{-1}\underline{x'v^{-1}y}$$

(Draper and Smith, 1966, p. 79).

## Statistical Tests and Measures of Fit

To check whether the fitted relationship, $\hat{y}$, is due to chance, we use an F test based on these sums of squares and test the null hypothesis:

$$H_o: \beta_i = 0$$

against the alternative

$$H_1: \text{ not all } \beta_i \text{ are equal to zero}$$

If $H_o$ is true, then

$$F = \frac{\sum(\hat{y}_i - \bar{y})^2/(m-1)}{\sum(y_i - \hat{y})^2/(n-m)}$$

has an F distribution with (m-1, n-m) degrees of freedom.

Another null hypothesis which can be tested is that a subset of the regression coefficients have specified values, for example, zero. Let the standard regression model be called the full model, FM, and the model with some parameters taking specified values be called the reduced model, RM. If the reduced model gives as good a fit as the full model, the null hypothesis is not rejected.

Let $y_i$ and $y_i^*$ be the predicted values of the full and reduced models, respectively, and suppose the reduced model has k distinct parameters. Then

$$SSE(FM) = \sum(y_i - y_i)^2$$

and

$$SSE(RM) = \sum(y_i - y_i^*)^2$$

represent the sums of squares due to error. The ratio

$$F = \frac{(SSE(RM) - SSE(FM))/(p+1-k)}{SSE(FM)/(n-p-1)}$$

has an F distribution with (p+1-k) and (n-p-1) degrees of freedom (Chatterjee and Price, 1977, p. 57). If F exceeds the tabled F value then the result is significant, i.e., the reduced model is not satisfactory and the null hypothesis, with its suggested values for some of the $\beta_i$ is rejected. Rao, 1973 and Searle, 1971, have proofs and derivations of the hypothesis testing methods. Violations of model assumptions may invalidate any formal statistical inference procedures relating to hypothesis testing or interpretation of the model.

The SPSS subprogram Regression output provides, in addition to the estimates $b_i$ needed to form the multiple regression equation, standardized regression coefficients and the standard error of estimate. The standardized coefficients are called beta weights, designated $beta_i$, and are calculated from the $b_i$ using the relationship

$$beta_i = b_i(s_{x_i}/s_y)$$

where $s_{x_i}$ is the standard deviation of $x_i$ and $s_y$ is the standard deviation of y. The beta weights have a standard deviation of one and a mean of zero. Using these weights we can compare the relative effect of y of each independent variable. The standard error of estimate (SEE) is the standard deviation of the residuals, i.e., the standard deviation of the predicted values from the observed values.

$$SEE = \sqrt{\frac{\sum(y - \hat{y})^2}{n - p - 1}}$$

SEE is thought of as the average error in predicting y using the regression equation and is used to evaluate the accuracy of the prediction.

$R^2$, the square of the multiple correlation coefficient, frequently called the coefficient of determination, can also be derived from the sums of squares.

$$R^2 = \frac{\text{SS due to regression}}{\text{SS about the mean}}$$

$$= \frac{\text{SS about the mean - SS of residuals}}{\text{SS about the mean}}$$

$$= 1 - \left[\sum(y_i - \hat{y}_i)^2 / \sum(y_i - \bar{y})^2\right]$$

$R^2$ is interpreted as the proportion of total variability which is explained by the regression equation. It has a range between zero and one. It can be thought of as a simple r between y and $\hat{y}$ since $\hat{y}$ can be considered as a single variable constructed from the regression equation (Kim and Kohout, 1975, p. 331). When the model is a good fit of the data, the observed and predicted values will be close to each other and $R^2$ will be close to one. If the linear model is a poor fit, the best predicted value for $y_i$ is $\bar{y}$ since in the absence of any relationship, the sample mean minimizes the sum of squared deviations. We can test the null hypothesis that the multiple correlation is zero. The test statistic used is

$$F = \frac{\text{explained SS}/p}{\text{unexplained SS}/(n-p-1)}$$

or,

$$F = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2/p}{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2/(n-p-1)}$$

where F is distributed approximately as an F distribution with (p, (n-p-1)) degrees of freedom (Kim and Kohout, 1975, p. 335). So, $R^2$ serves as a summary measure of goodness of fit, though a large $R^2$ does not imply that model assumptions have been met. This requires an analysis of the residuals.

Forward (Stepwise) Inclusion

Variables are entered into the regression equation one at a time in a procedure known as forward (stepwise) regression. The first independent variable inserted into the equation is the one that will cause the largest $R^2$, i.e., it explains the greatest amount of variance in the dependent variable and will cause the greatest reduction in the residual sum of squares (Mosteller and Tukey, 1977, p. 388). The variable which explains the largest amount of variance when combined with the first will be entered second, and so on. At each step, the variable which adds the most to the multiple correlation is entered in the equation, yielding the best k-predictor equation among those equations which contain the first k-1 variables selected. At each step, an F ratio is calculated for the variables not in the equation. This F ratio is the F value which would be obtained if the variable were entered into the equation on the next step. The process of adding variables stops either when the F ratio reaches .01 or when the tolerance level reaches .001, where tolerance is defined as the proportion of variance of the variable not already explained by the variables previously entered into the equation (Kim and Kohout, 1975, p. 346).

# CHAPTER IV

# METHOD

## Overview

Files of students who graduated from the Graduate School of Business over a two year period were studied to determine whether an accurate equation could be developed to predict a student's graduate grade point average. The students were classified into two groups, male and female, and each group was analyzed separately. Multiple regression equations were generated by computer to form the prediction equations. The standardized residuals were plotted against the predicted values and each independent variable in the equations. These plots were then analyzed for possible model violations. Several procedures were tried in the attempt to correct for the existing violations and to improve the accuracy of the prediction. A new regression equation was formulated at each step. These procedures include the elimination of outliers, the addition of interaction terms, the addition of squared variable terms, the addition of dummy variables and the use of a transformation (weighted least squares). The standardized residuals were examined at several stages to insure that existing violations were being corrected for before a final equation was selected as the optimal one for each group.

## The Subjects and the Variables

The data for this study were collected over a two-year period from the records of the classes which graduated from the Loyola Graduate School of Business in January 1979, May 1979, January 1980 and May 1980. These students came from 127 undergraduate colleges and universities and waited an average of three years after graduation before beginning their Masters of Business Administration program. There were two hundred eighty males and one hundred twenty females. The mean entering age for males was twenty-six, for females, twenty-five.

Eight independent variables were used in the initial regression analysis. They were age upon entering the Graduate School of Business, number of undergraduate institutions attended, number of years between completion of the undergraduate degree and beginning graduate studies, number of months of full-time work experience at the beginning of graduate studies, scores on the verbal and quantitative sections of the Graduate Management Admissions Test (GMAT), undergraduate grade point average and an undergraduate school quality index. The dependent variable was the overall graduate grade point average for each student of courses taken at the Loyola Graduate School of Business. Each variable was assigned a four-letter abbreviation for ease in referring to them and in formulating later interaction terms and squared variable terms.

Table 2 lists these variables and the abbreviations which are used to refer to them in the text of this dissertation. Table 3 shows the means and standard deviations for each group on the nine variables. Tables 4 and 5 list the correlations between these variables for the female and male samples, respectively.

The GMAT is a standardized exam required both by the Loyola Graduate School of Business and the American Assembly of Collegiate Schools of Business (the accrediting body) for applicants to the Graduate School of Business. Schrader (1979, p. 15) reports the median validity coefficients for an optimally weighted total of undergraduate average grades in combination with GMAT test scores. These are shown in Table 6. He also reports (p. 16) the reliability coefficients for the four forms of the GMAT. These are shown in Table 7.

From the undergraduate school attended, it was possible to obtain a measure of undergraduate school quality based on the _Statistical Summary by Undergraduate College Attended_, published by the Educational Testing Service. This quality index is the mean total GMAT score of all students taking the GMAT from each college for the years 1971-1976. Due to the confidential nature of this information, it is not reproduced here. However, the list of undergraduate schools from which students in this sample graduated may be found in Table 8.

Table 2. Regression Variables and Their Abbreviations.

| Variable | Abbreviation |
|---|---|
| Dependent variable: | |
| Graduate grade point average | GGPA |
| Independent variables: | |
| Age when entering the graduate school of business | AGE |
| Number of undergraduate institutions attended | NOSC |
| Number of years between completion of undergraduate degree and entering the graduate school of business | YRSB |
| Number of months of full-time work experience at the beginning of graduate work | WORK |
| Score on the verbal section of the GMAT | VERB |
| Score on the quantitative section of the GMAT | QUAN |
| Undergraduate grade point average | UGPA |
| Undergraduate school quality index | SCAV |

Table 3. Means and Standard Deviations of Females and Males on Each of the Nine Variables.

Male Group (n = 280):

| Variable | Mean | Standard Deviation |
|---|---|---|
| GGPA | 2.8020 | 0.3955 |
| AGE | 26.1107 | 4.2085 |
| NOSC | 1.5464 | 0.7653 |
| YRSB | 3.2036 | 3.4211 |
| WORK | 42.4179 | 36.9204 |
| VERB | 30.4071 | 5.9657 |
| QUAN | 30.2964 | 6.0311 |
| SCAV | 478.5393 | 34.8719 |
| UGPA | 2.9303 | 0.3598 |

Female Group (n = 120):

| Variable | Mean | Standard Deviation |
|---|---|---|
| GGPA | 2.7624 | 0.3723 |
| AGE | 25.4500 | 4.3903 |
| NOSC | 1.7667 | 0.9234 |
| YRSB | 3.1333 | 3.8500 |
| WORK | 37.8667 | 35.9916 |
| VERB | 31.3083 | 5.0937 |
| QUAN | 26.8333 | 4.9490 |
| SCAV | 482.0000 | 33.6015 |
| UGPA | 3.0828 | 0.3382 |

Table 4. Correlation Between the Variables, Female Sample.

| | GGPA | AGE | NOSC | YRSB | WORK |
|---|---|---|---|---|---|
| GGPA | 1.00000 | 0.07762 | -0.03867 | 0.14791 | 0.13788 |
| AGE | 0.07762 | 1.00000 | 0.09038 | 0.79437 | 0.87931 |
| NOSC | -0.03867 | 0.09038 | 1.00000 | -0.04554 | 0.14369 |
| YRSB | 0.14791 | 0.79437 | -0.04554 | 1.00000 | 0.81131 |
| WORK | 0.13788 | 0.87931 | 0.14369 | 0.81131 | 1.00000 |
| VERB | 0.03350 | 0.24588 | -0.00959 | 0.23399 | 0.28611 |
| QUAN | 0.18806 | 0.07194 | 0.14588 | 0.03425 | 0.05502 |
| SCAV | 0.02075 | -0.06938 | -0.02085 | -0.01501 | -0.10877 |
| UGPA | 0.18196 | -0.16699 | 0.10170 | -0.19199 | -0.18917 |

| | VERB | QUAN | SCAV | UGPA |
|---|---|---|---|---|
| GGPA | 0.03350 | 0.18806 | 0.02075 | 0.18196 |
| AGE | 0.24588 | 0.07194 | -0.06938 | -0.16699 |
| NOSC | -0.00959 | 0.14588 | -0.02085 | 0.10170 |
| YRSB | 0.23399 | 0.03425 | -0.01501 | -0.19199 |
| WORK | 0.28611 | 0.05502 | -0.10877 | -0.18917 |
| VERB | 1.00000 | 0.11939 | 0.14582 | 0.02817 |
| QUAN | 0.11939 | 1.00000 | 0.04164 | 0.00797 |
| SCAV | 0.14582 | 0.04164 | 1.00000 | -0.13158 |
| UGPA | 0.02817 | 0.00797 | -0.13158 | 1.00000 |

Table 5. Correlations Between the Variables, Male Sample.

| | GGPA | AGE | NOSC | YRSB | WORK | VERB | QUAN | SCAV | UGPA |
|---|---|---|---|---|---|---|---|---|---|
| GGPA | 1.00000 | 0.14862 | -0.04750 | 0.24101 | 0.14848 | 0.19489 | 0.29283 | 0.11967 | 0.19200 |
| AGE | 0.14862 | 1.00000 | 0.15810 | 0.72609 | 0.86120 | 0.07243 | 0.09614 | -0.03829 | -0.10686 |
| NOSC | -0.04750 | 0.15810 | 1.00000 | -0.04538 | 0.16150 | 0.00684 | -0.02124 | -0.16137 | -0.02010 |
| YRSB | 0.24101 | 0.72609 | -0.04538 | 1.00000 | 0.67009 | 0.10990 | 0.08218 | -0.06026 | -0.15107 |
| WORK | 0.14848 | 0.86120 | 0.16150 | 0.67009 | 1.00000 | 0.12641 | 0.11421 | -0.06413 | -0.22725 |
| VERB | 0.19489 | 0.07243 | 0.00684 | 0.10990 | 0.12641 | 1.00000 | 0.20862 | 0.10907 | -0.12116 |
| QUAN | 0.29283 | 0.09614 | -0.02124 | 0.08218 | 0.11421 | 0.20862 | 1.00000 | 0.25772 | -0.12608 |
| SCAV | 0.11967 | -0.03829 | -0.16137 | -0.06026 | -0.06413 | 0.10907 | 0.25772 | 1.00000 | -0.15176 |
| UGPA | 0.19200 | -0.10686 | -0.02010 | -0.15107 | -0.22725 | -0.12116 | -0.12608 | -0.15176 | 1.00000 |

Table 6. Median validity coefficients for undergraduate average grades in combination with GMAT scores.

| Years in which studies were done | Number of schools | Correlation of first-year grades with undergraduate grades and GMAT acores |
|---|---|---|
| 1963-64 | 17 | .47 |
| 1967-70 | 17 | .46 |
| 1967-70 | 67 | .39 |
| 1977-78 | 10 | .45 |
| 1978-79 | 25 | .48 |

Table 7. Reliability coefficients for the four forms of the GMAT.

| Form | Reliability coefficient of: | | |
|---|---|---|---|
| | GMAT Total | Verbal | Quantitative |
| A | .92 | .90 | .86 |
| B | .92 | .89 | .87 |
| C | .92 | .90 | .88 |
| D | .93 | .90 | .88 |

Table 8. Undergraduate Schools Represented in the Sample.

Arizona State University
Ashland College
Auburn University
Augustana College
Barat College
Beloit College
Blackburn College
Boston University
Bradley University
California State University, Fresno
California State University, Fullerton
Calumet College
Chicago State University
College of St. Thomas
Colorado State University
Colorado University
Cornell University
C W Post College of Long Island
Depaul University
Drake University
Eastern Illinois University
Elmhurst College
Fairfield University
Fresno State College
Georgetown University
George Williams College
Georgia Institute of Technology
Governors State University
Hope College
Illinois Benedictine College
Illinois Institute of Technology
Illinois State University
Illinois Wesleyan University
Indiana University
Iowa State University
John Carroll University
Kansas State University
Kendall College
Kent State
LeMoyne College
Lake Forest College
Lewis University
Loras College
Loyola University of Chicago
Manhattan College
Marquette University
Maryville College
Massachusetts College of Pharmacy
Massachusetts Institute of Technology
Miami University
Michigan State University
Milwaukee School of Engineering
Monmouth College
Montana State University
Moorhead State University
Mount Mary College
Mundelein College
New York Institute of Technology
North Central College
North Park College
Northeastern Illinois University
Northern Illinois University
Northwestern University
Ohio State University
Pennsylvania State University
Providence College
Purdue University, Hammond
Purdue University, Lafayette
Quincy College
Randoply-Macon College
Regis College
Rensselaur Polytechnic Institute
Ripon College
Roosevelt University
Rutgers State University
San Jose State University
Scripps College
Seton Hall University
Southern Illinois University
Spring Hill College

(Table 8, continued)

St Ambrose College
St Francis College
St Josephs College
St Louis University
St Marys College
St Norbert College
St Olaf College
Stevens Institute of Technology
SUNY Center at Buffalo
Texas Womans University
Theil College
Tri State University
Trinity College
Trinity University
Union College
University of Chicago
University of Colorado
University of Dayton
University of Denver
University of Detroit
University of Florida
University of Georgia
University of Illinois, Chicago
University of Illinois, Medical Center
University of Illinois, Urbana
University of Iowa
University of Maryland
University of Miami
University of Michigan
University of Minnesota
University of Missouri
University of Notre Dame
University of Oklahoma
University of San Diego
University of Texas
University of the Americas
University of Toledo
University of Western Ontario
University of Wisconsin, Madison
University of Wisconsin, Stout
University of Wisconsin, Whitewater
Valparaiso University
Vassar College
Virginia Polytechnic Institute and State University
Wellesley College
Western Illinois University
Western Michigan University
Winona State University

## The Regression Procedures

The process of developing the prediction equation for females began with an initial regression using GGPA as the dependent variable and the eight independent variables previously listed in Table 2. The standardized residuals were plotted and analyzed. Several outliers were eliminated and a new regression equation was calculated. The standardized residuals were again graphed and reviewed. Based on the results of the regression and several graphs, additional terms were added to the equation in an attempt to correct the model violations appearing in the plots and improve the value of $R^2$. Finally, a transformation was applied and new regression parameters were calculated. Standardized residuals were again plotted and checked for model violations.

For the male sample, the first regression was run using GGPA as the dependent variable and the eight independent variables of Table 2. The standardized residuals were graphed and analyzed. Several outliers were removed. A second regression was run and the standardized residuals were again plotted and analyzed. Following this regression, several other regressions were run with various additional terms included in each run. Finally, a regression was run using the best terms from the previous equations. The standardized residuals for this last regression were plotted and reviewed.

For each group, the initial and final regressions were compared with respect to fit and shape of the standardized residuals plots.

# CHAPTER V

# THE RESULTS

## Introduction

The sample was separated into two groups on the basis of sex, and each group was analyzed individually. A sequence of multiple regression programs were run on each group and the results from each program were recorded. Outliers were eliminated from both groups and the plots of standardized residuals versus the fitted values and versus the independent variables were analyzed at several stages. The final equation for females used a transformation to achieve optimum predictive accuracy. As a final equation for males, those variables which performed best, in terms of the amount contributed to $R^2$ and the significance of the F value for each parameter, were combined and a regression was run using them to obtain the best equation using the least number of variables.

Fitted equations which will be referred to at various places in the text are assigned numbers when they first appear and thereafter are referred to by their number. Fitted values, as opposed to observed values, are denoted by a hat ($\wedge$) above the variable or value name.

Regression and Residual Plots for the Female Sample

The first multiple regression for females used GGPA as the dependent variable and AGE, NOSC, YRSB, WORK, VERB, QUAN, SCAV and UGPA as independent variables. The regression was run in a stepwise fashion using the Statistical Package for the Social Sciences (SPSS) subprogram Regression (Kim and Kohout, 1975). The results from the run are shown in Table 9. The Durbin-Watson statistic was calculated to be 1.91383, indicating that autocorrelation was not present.

The final $R^2$ for this run was .13548, implying that a little more than 13% of the variance in GGPA was accounted for by variance in the independent variables. The F ratio used to test the significance of $R^2$ has a value of 2.17446 with (8, 111) degrees of freedom. This is just significantly different from zero at the .05 level, but not at the .01 level (Hays, 1973, p. 888). QUAN, UGPA, and YRSB each contributed a little over 3% to $R^2$, AGE contributed 1.6%, the remaining variables, NOSC, WORK, SCAV, and VERB each contributed less than 1%.

The prediction equation generated by this program is

$$\widehat{GGPA} = .0154QUAN + .2758UGPA + .0135YRSB - .0454NOSC + .0037WORK - .0247AGE + .0008\ SCAV - .0055VERB + 1.7937. \tag{1}$$

Table 9. Initial Regression, Female Sample.

| VARIABLE | MULTIPLE R | R SQUARE | B | BETA |
|---|---|---|---|---|
| QUAN | 0.18806 | 0.03537 | 0.015386 | 0.20450 |
| UGPA | 0.26064 | 0.06793 | 0.275832 | 0.25052 |
| YRSB | 0.31654 | 0.10020 | 0.013524 | 0.13983 |
| NOSC | 0.32645 | 0.10657 | -0.045370 | -0.11252 |
| WORK | 0.33244 | 0.11052 | 0.003748 | 0.36233 |
| AGE | 0.35673 | 0.12726 | -0.024694 | -0.29117 |
| SCAV | 0.36151 | 0.13069 | 0.000832 | 0.07506 |
| VERB | 0.36808 | 0.13548 | -0.005467 | -0.07479 |
| (CONSTANT) | | | 1.793780 | |

| | |
|---|---|
| MULTIPLE R | 0.36808 |
| R SQUARE | 0.13548 |
| STANDARD ERROR | 0.35846 |

| ANALYSIS OF VARIANCE | DF | SUM OF SQUARES | MEAN SQUARE | F |
|---|---|---|---|---|
| REGRESSION | 8 | 2.23526 | 0.27941 | 2.17446 |
| RESIDUAL | 111 | 14.26294 | 0.12849 | |

The standard error of prediction is .35846. Three variables, NOSC, AGE and VERB were assigned negative weights. This indicates they are acting as suppressor variables. They have low correlations with GGPA but high correlations with some of the other independent variables in the equation. Variables with negative weights act to eliminate irrelevant variance in the other independent variables (Anastasi, 1976, p. 183). When this variance is removed, the predictive power of the other variables is increased (Nunnally, 1967, p. 162).

Residuals were calculated by subtracting the observed GGPA from the fitted GGPA. The residuals were standardized using the SPSS formula

$$\text{standardized residual} = \frac{\text{residual}}{\text{standard error of regression}}$$

and plotted against GGPA and each of the independent variables in order to look for violations of the model assumptions. The residual plots are shown in Figures 13 through 21. The first portion of the analysis of the residuals plots is the removal of the extreme data points, or outliers. The points which are to be removed are shown as circled data points in the plots. The cards generating those data points were removed and do not occur in any of the remaining analyses.

Figure 13. Graph of standardized residuals versus $\hat{y}$, initial regression, female sample.

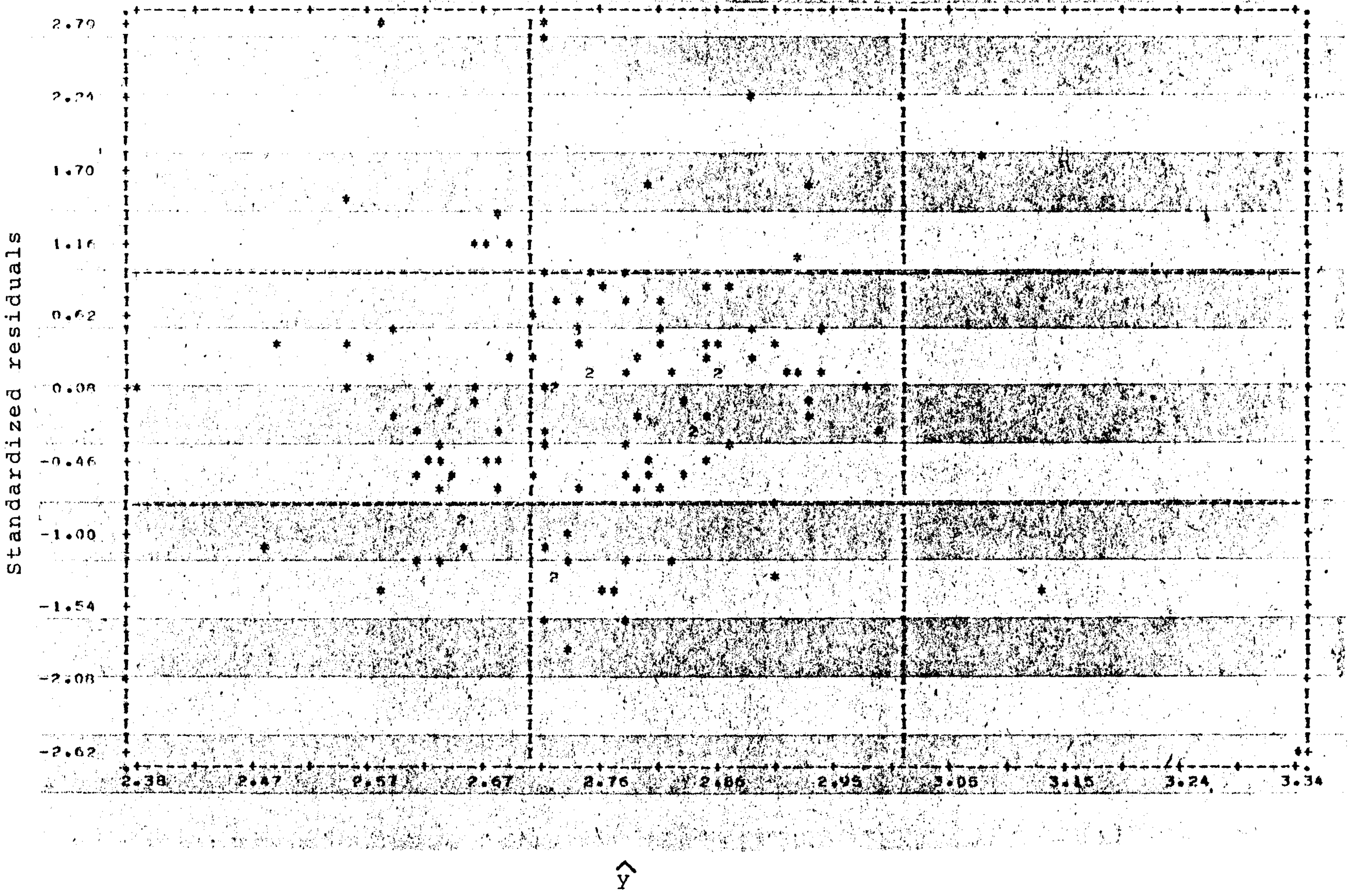


Figure 14. Graph of standardized residuals versus QUAN, initial regression, female sample.

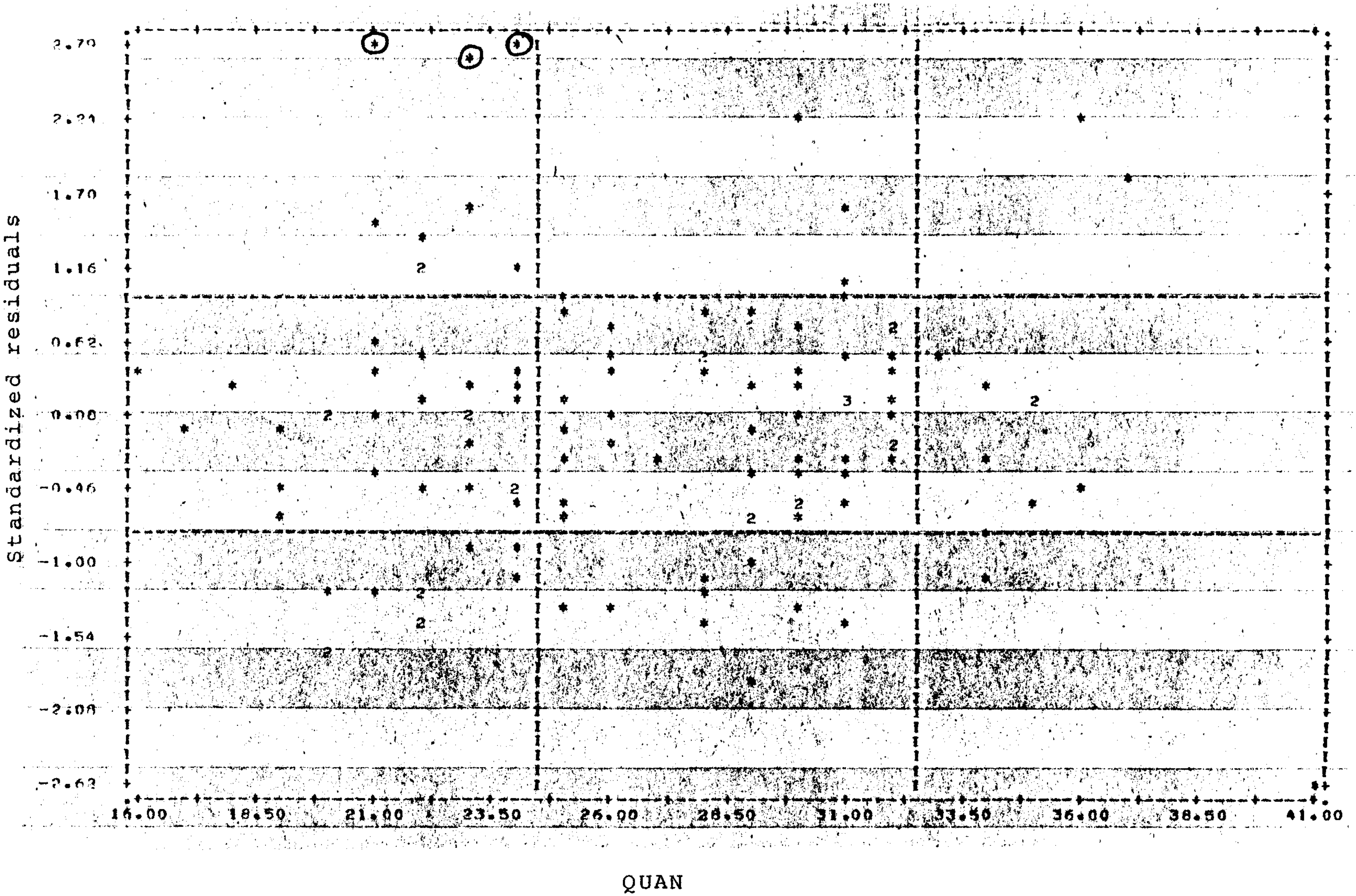


Figure 15. Graph of standardized residuals versus UGPA, initial regression, female sample.

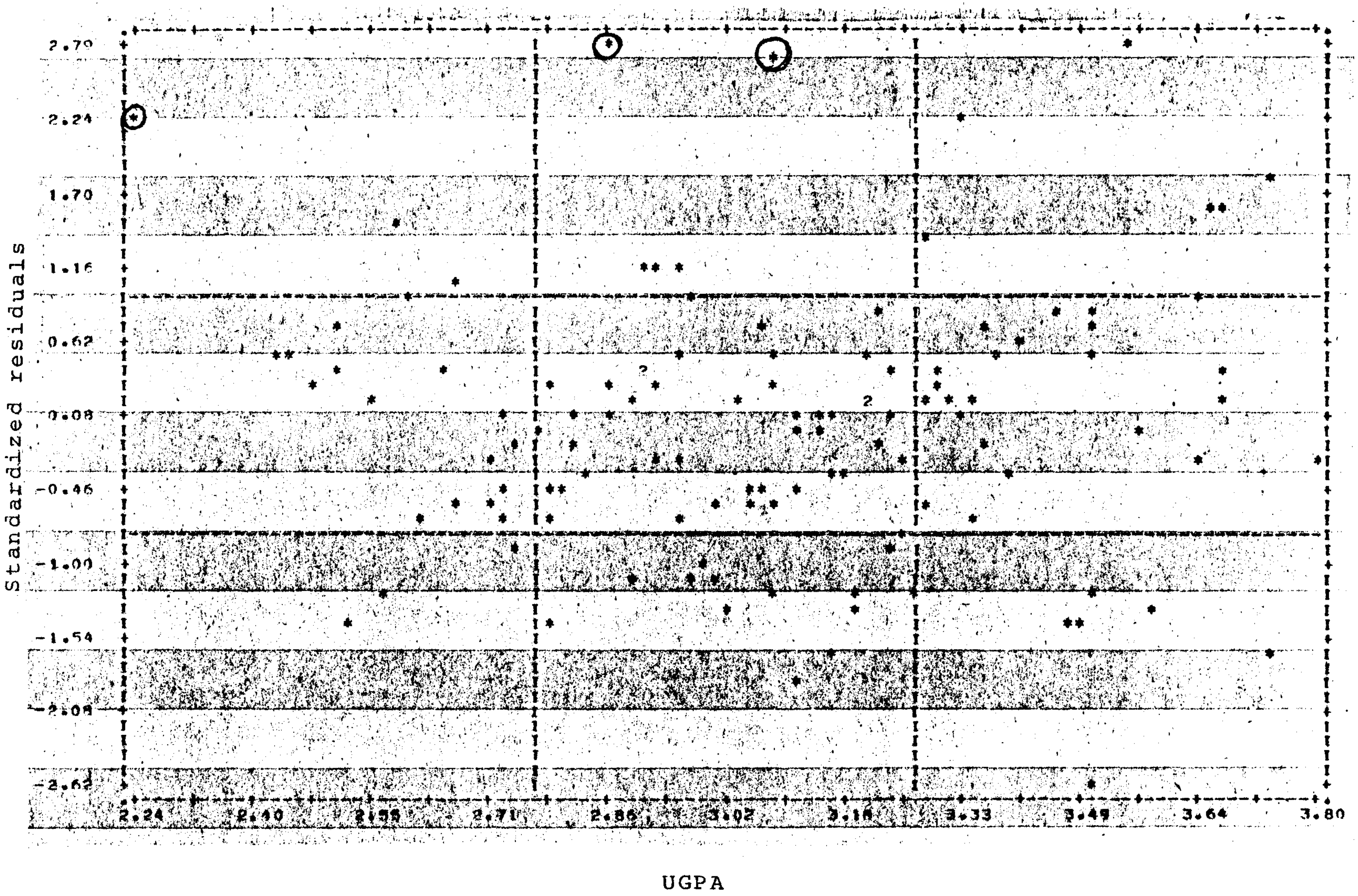


Figure 16. Graph of standardized residuals versus YRSB, initial regression, female sample.

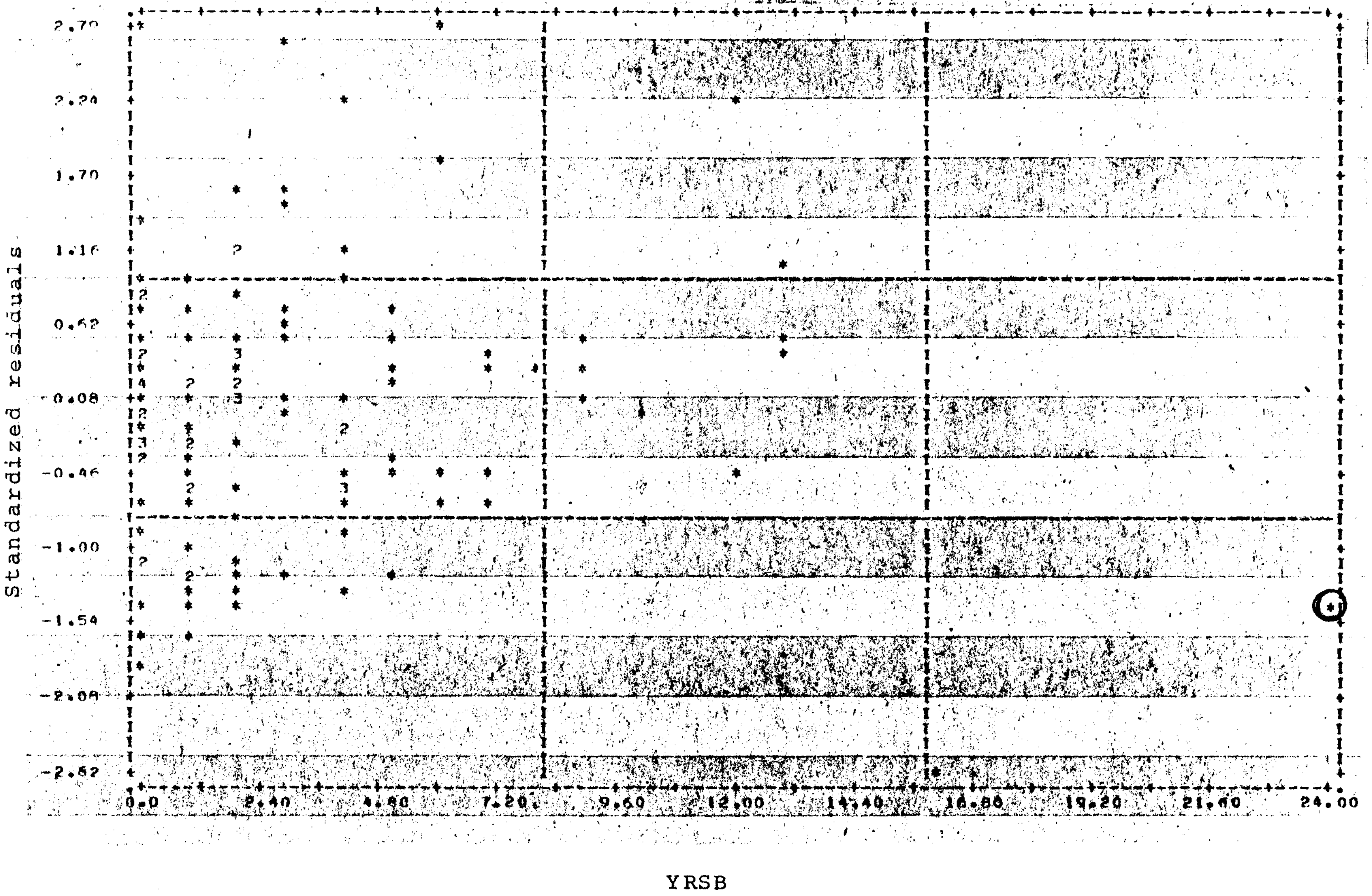


Figure 17. Graph of standardized residuals versus NOSC, initial regression, female sample.

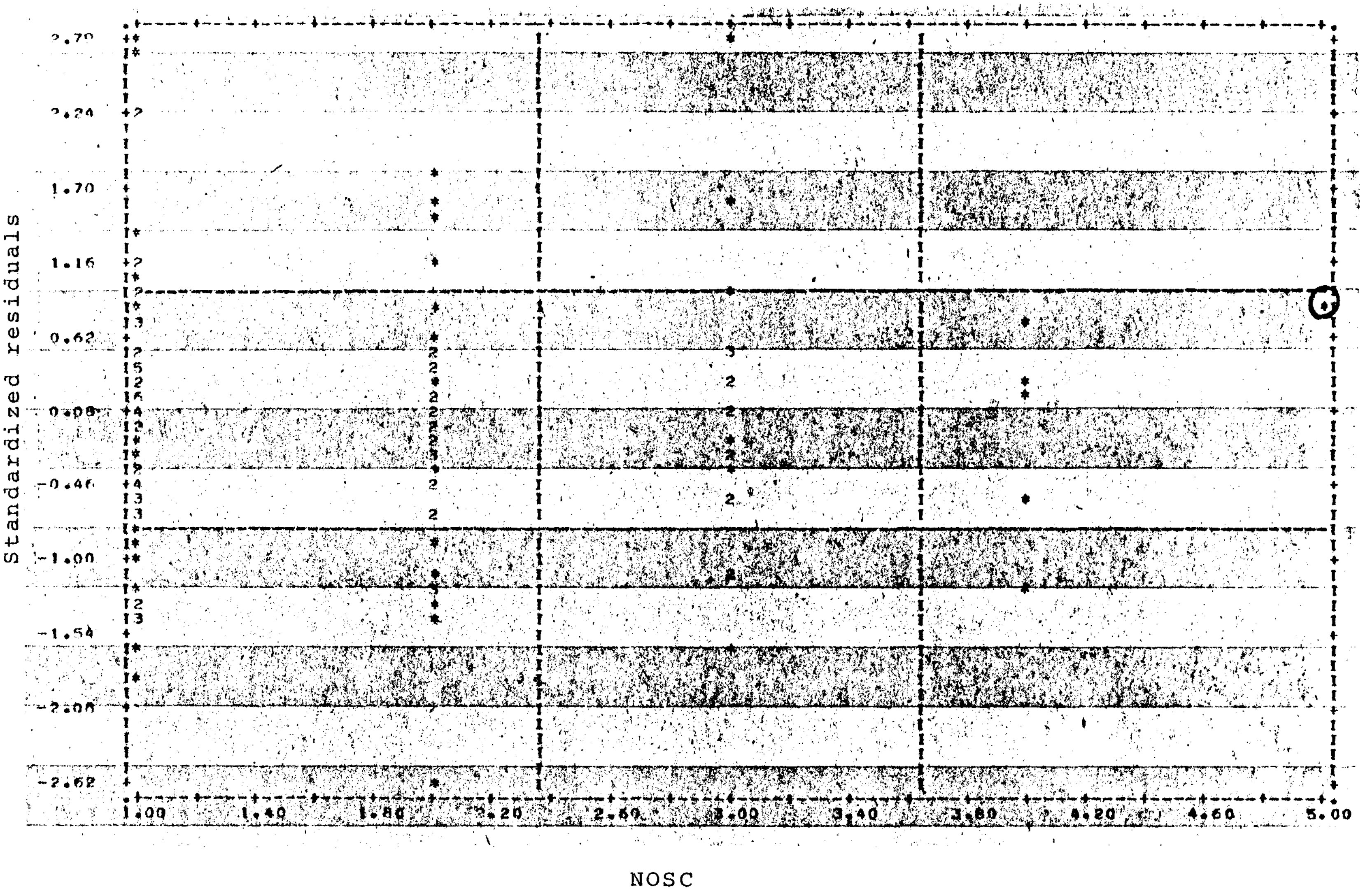


Figure 18. Graph of standardized residuals versus WORK, initial regression, female sample.

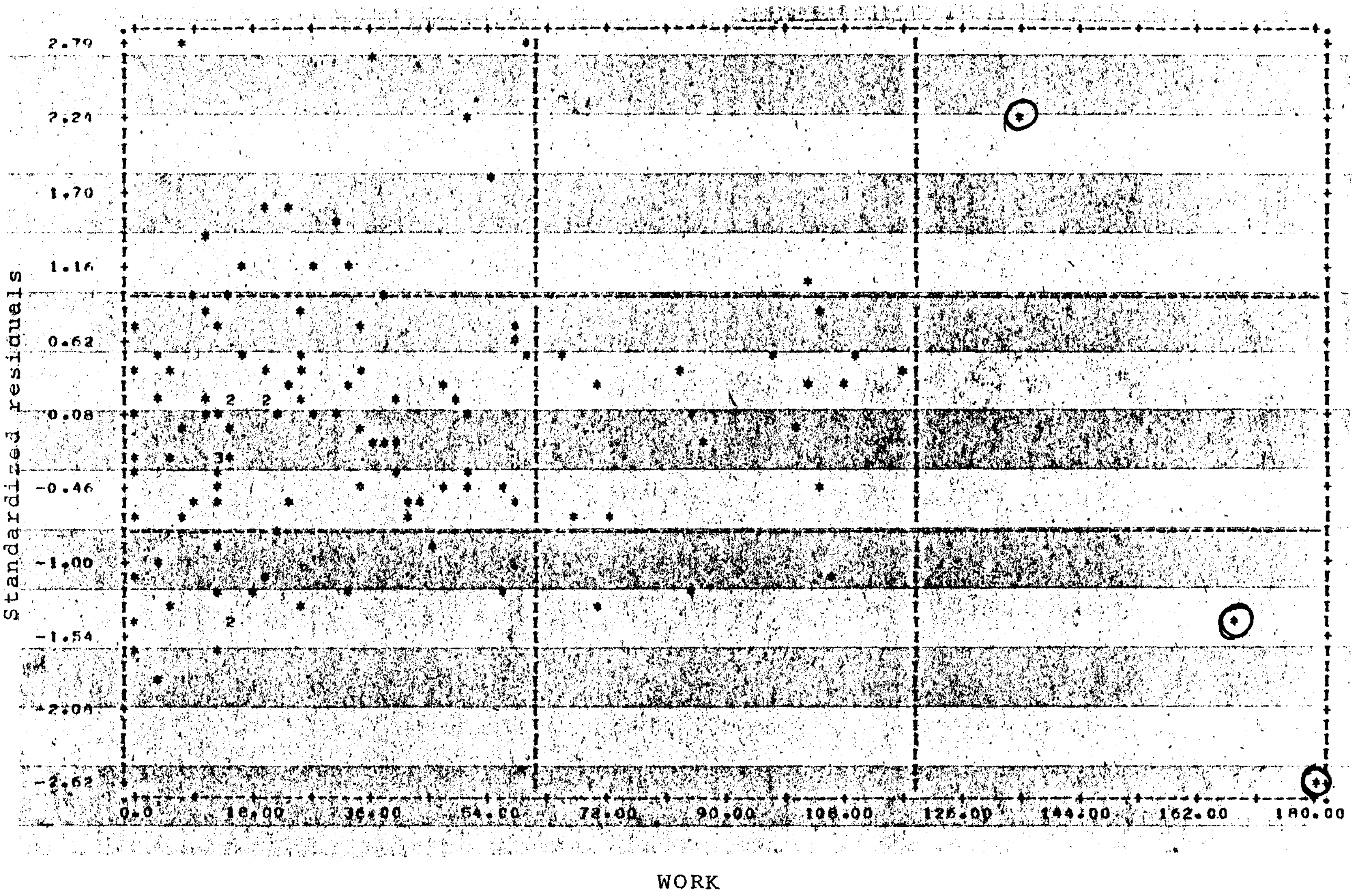


Figure 19. Graph of standardized residuals versus AGE, initial regression, female sample.

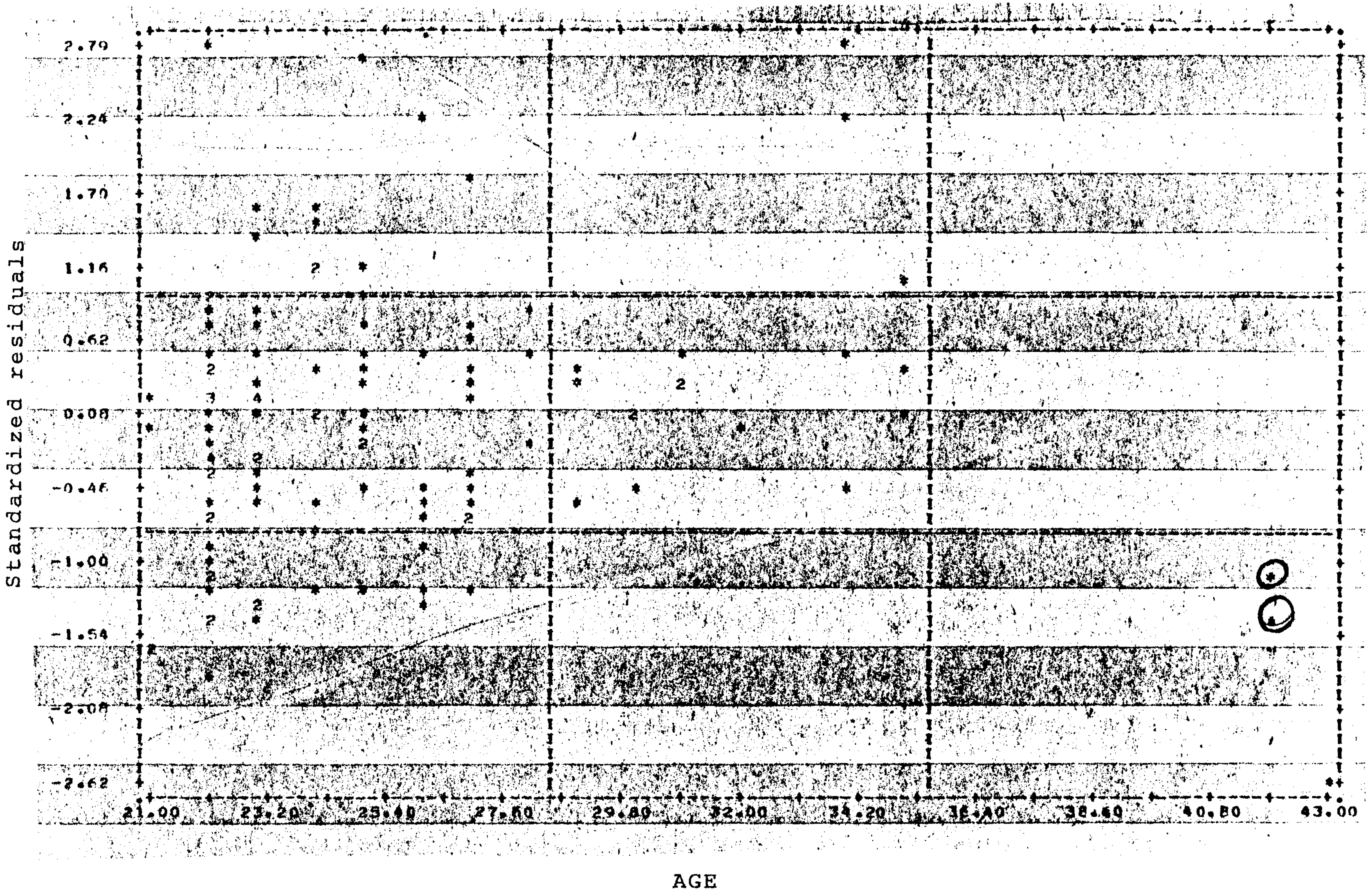


Figure 20. Graph of standardized residuals versus SCAV, initial regression, female sample.

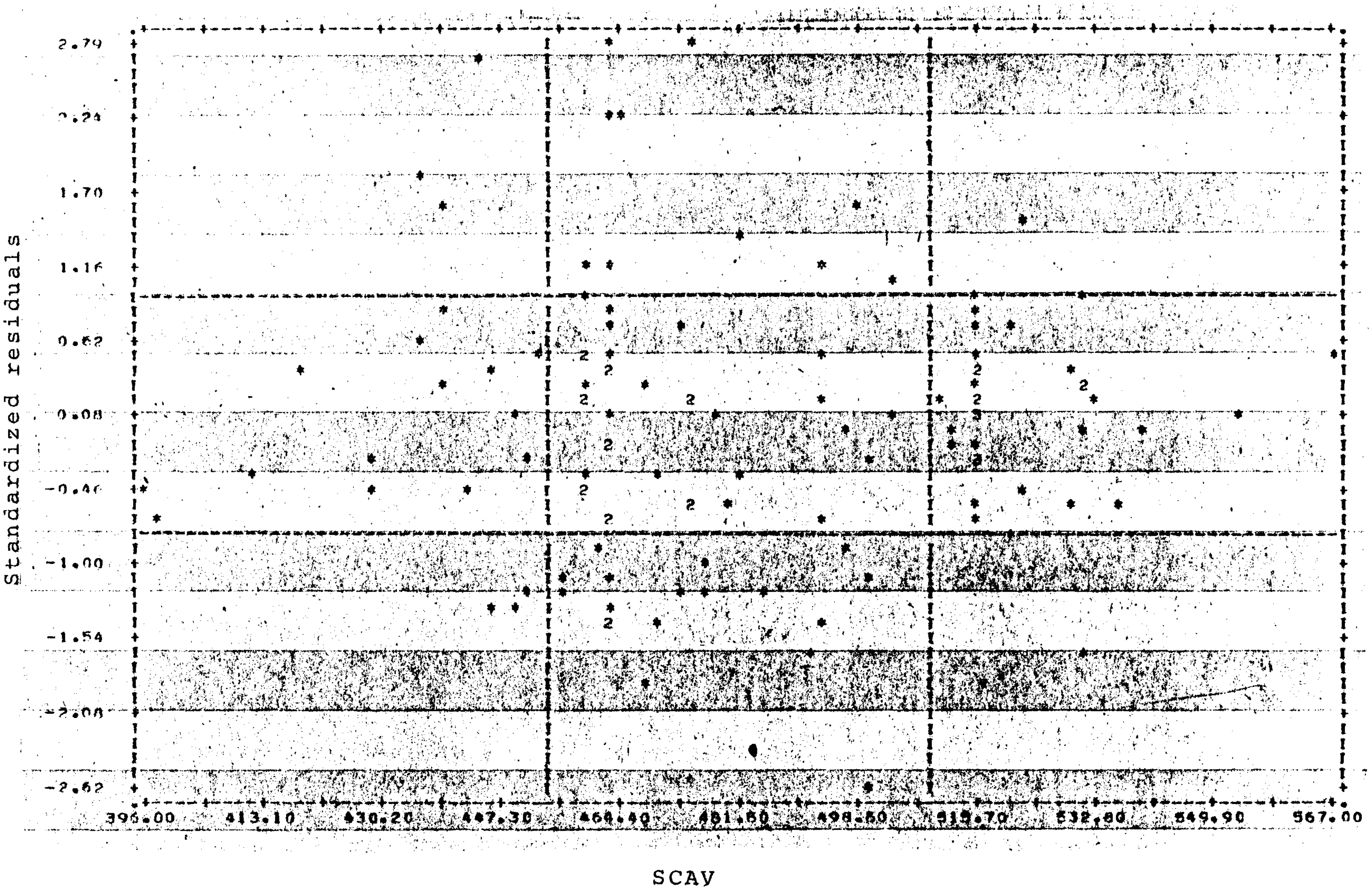
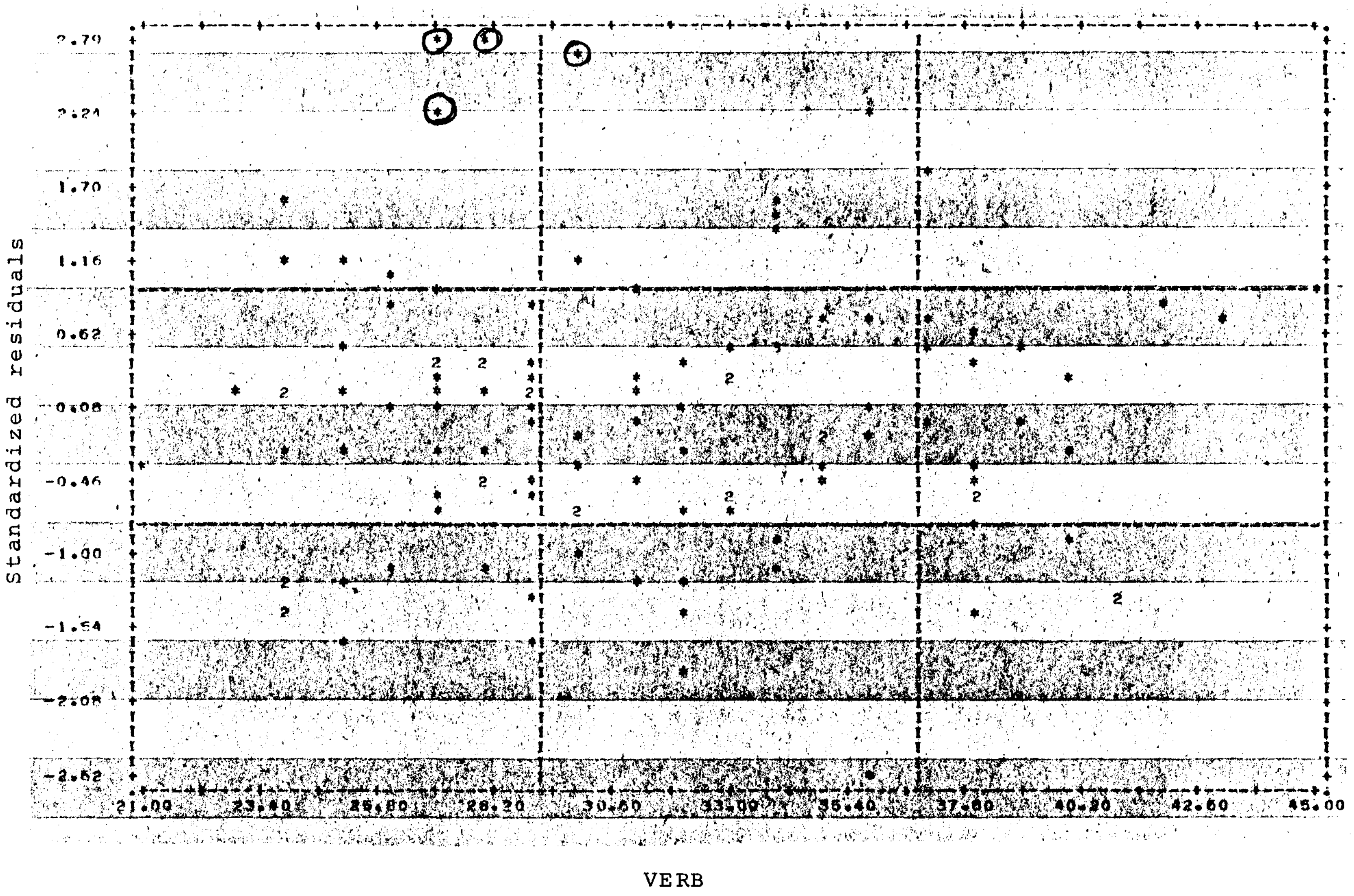


Figure 21. Graph of standardized residuals versus VERB, initial regression, female sample.



A total of eleven cards were removed, reducing the female sample to 109 cases. A second regression was run on the reduced sample. The results are shown in Table 10. The $R^2$ increased to .342, or 34%. The standard error was reduced to .266. The prediction equation generated by this regression was

$$(2) \quad \widehat{GGPA} = .0244QUAN + .4164UGPA + .0142YRSB + .0018SCAV - .0719NOSC + .0037WORK - .0031VERB - .0043AGE + .0542.$$

Comparing this with equation (1), the first prediction equation, we see that the size of the coefficients changed for all predictors except WORK, illustrating the sensitivity of the parameters to outliers. Suppose we hypothesize a student with the following scores on the variables:

| | |
|---|---|
| QUAN | 26 |
| VERB | 31 |
| UGPA | 2.80 |
| SCAV | 460 |
| YRSB | 3 |
| NOSC | 2 |
| AGE | 23 |
| WORK | 30 |

Then, equation (1) would predict a GGPA of 2.66, while equation (2) would predict 2.50.

With outliers removed, standardized residuals were again calculated and plotted against $\hat{y}$ and each independent variable. The results are shown in Figures 22 through 30. The plots of the standardized residuals versus $\hat{y}$, QUAN,

Table 10. Regression with Outliers Removed, Female Sample.

| VARIABLE | MULTIPLE R | R SQUARE | B | BETA |
|---|---|---|---|---|
| QUAN | 0.30340 | 0.09205 | 0.024364 | 0.37003 |
| UGPA | 0.39558 | 0.15648 | 0.416416 | 0.43129 |
| YRSB | 0.51297 | 0.26314 | 0.014233 | 0.13933 |
| SCAV | 0.43728 | 0.28867 | 0.001849 | 0.20050 |
| NOSC | 0.55767 | 0.31099 | -0.071875 | -0.20160 |
| WORK | 0.58248 | 0.33928 | 0.003739 | 0.35020 |
| VERB | 0.58440 | 0.34152 | -0.003119 | -0.05095 |
| AGE | 0.58477 | 0.34196 | -0.004277 | -0.04566 |
| (CONSTANT) | | | 0.054171 | |

| | |
|---|---|
| MULTIPLE R | 0.58477 |
| R SQUARE | 0.34196 |
| STANDARD ERROR | 0.26633 |

| ANALYSIS OF VARIANCE | DF | SUM OF SQUARES | MEAN SQUARE | F |
|---|---|---|---|---|
| REGRESSION | 8 | 3.68604 | 0.46075 | 6.49571 |
| RESIDUAL | 100 | 7.09322 | 0.07093 | |

Figure 22. Graph of standardized residuals versus $\hat{Y}$, regression with outliers removed, female sample.

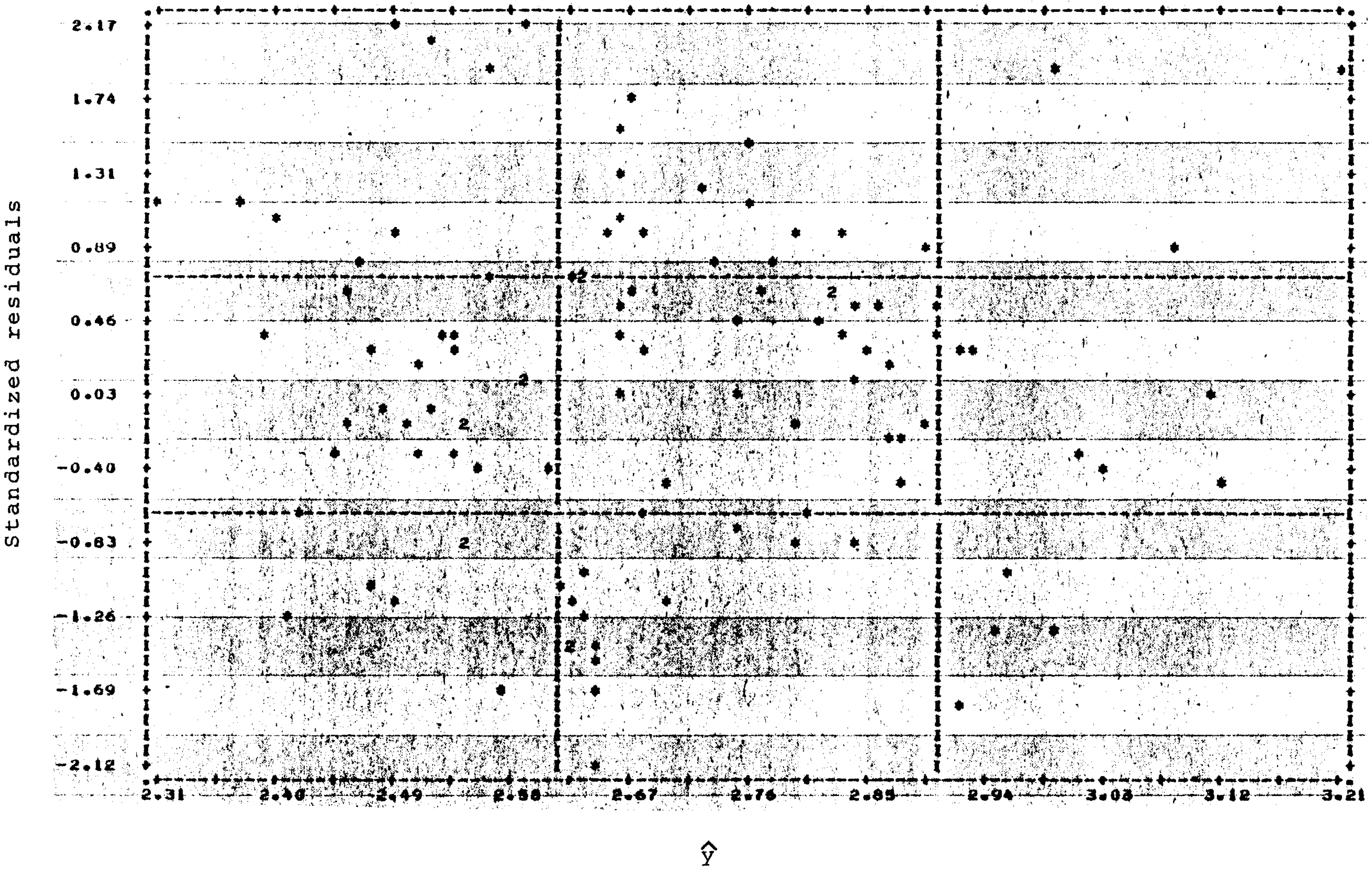


Figure 23. Graph of standardized residuals versus QUAN, regression with outliers removed, female sample.

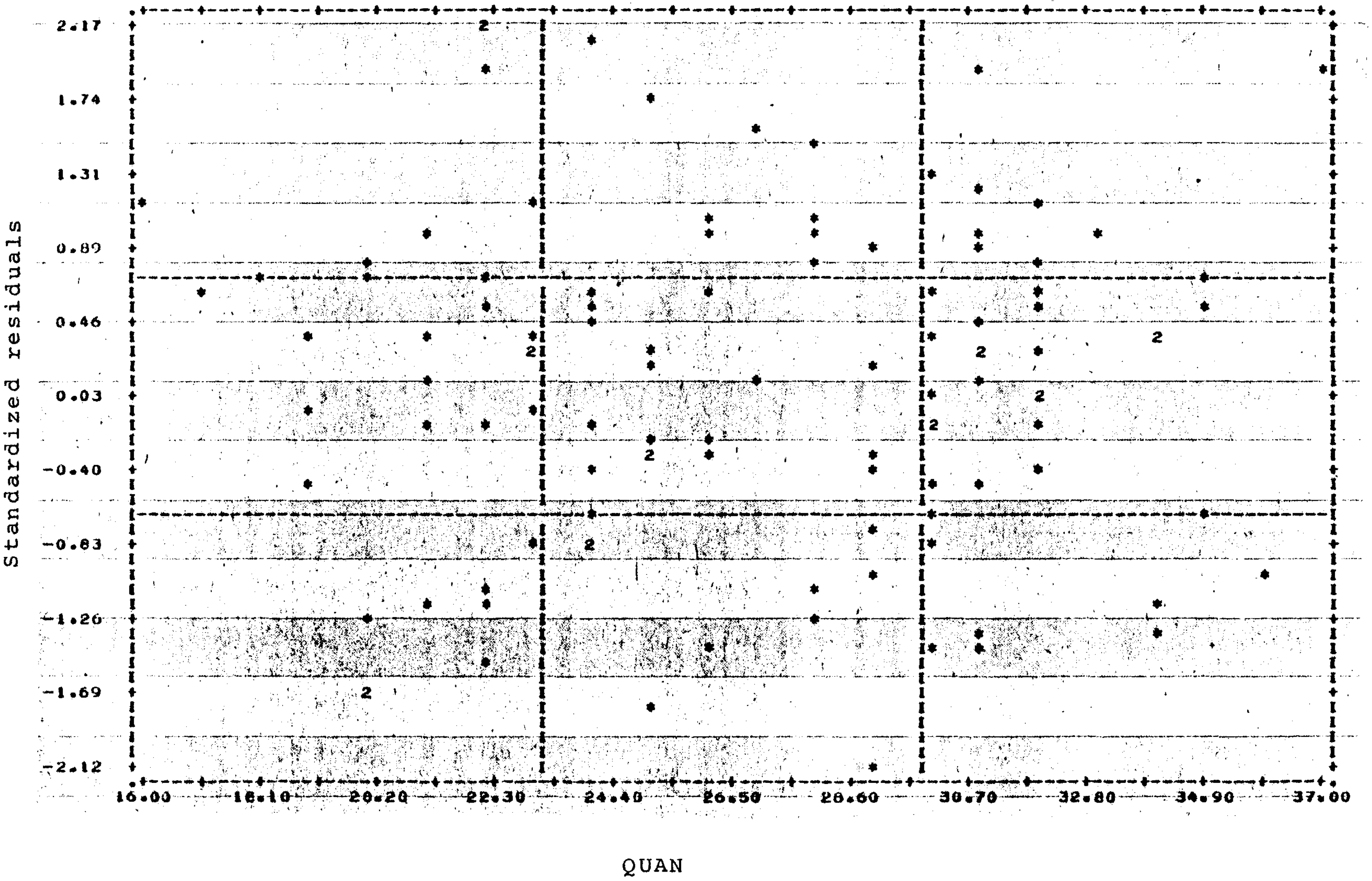


Figure 24. Graph of standardized residuals versus UGPA, regression with outliers removed, female sample.

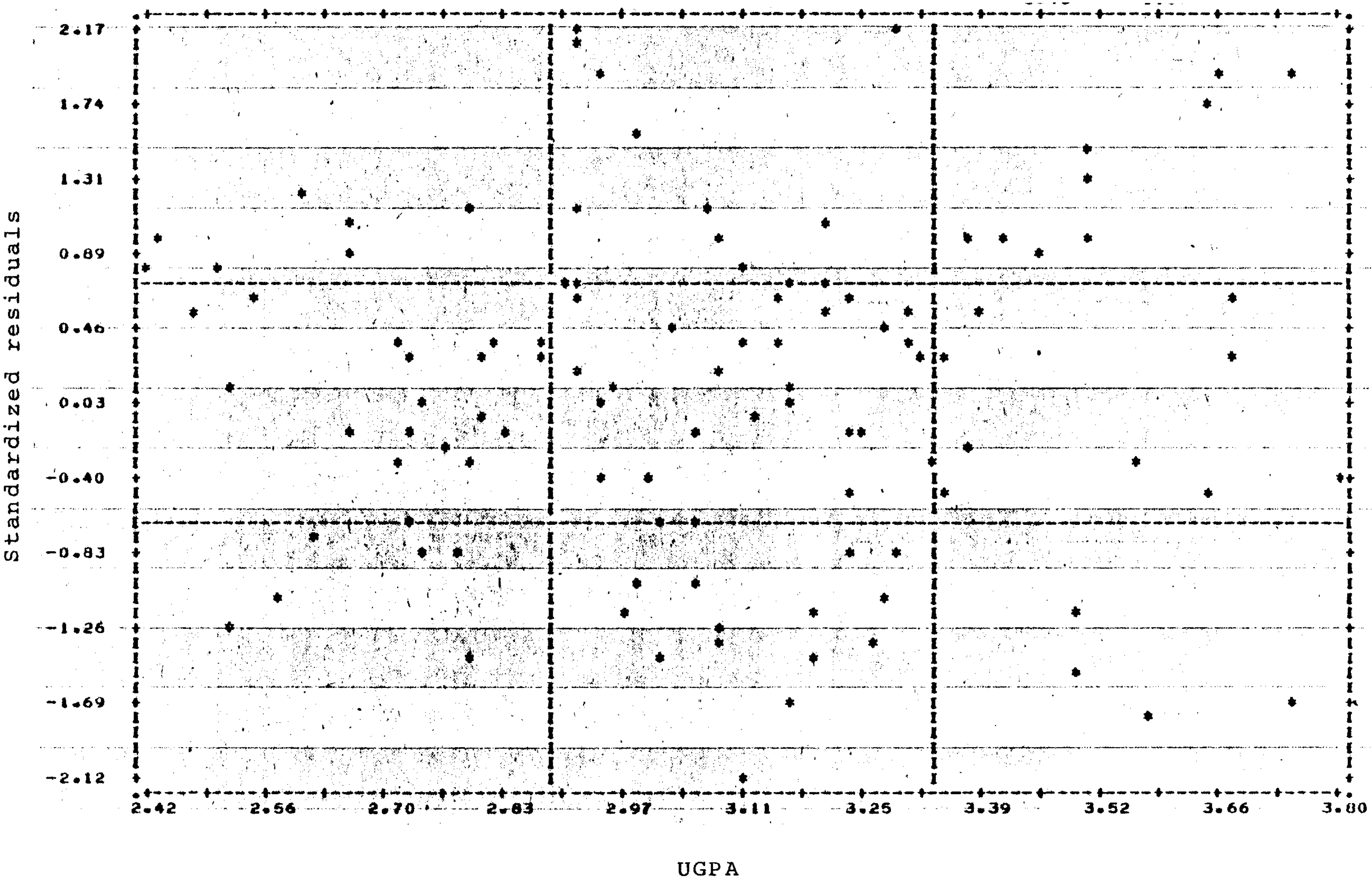


Figure 25. Graph of standardized residuals versus YRSB, regression with outliers removed, female sample.

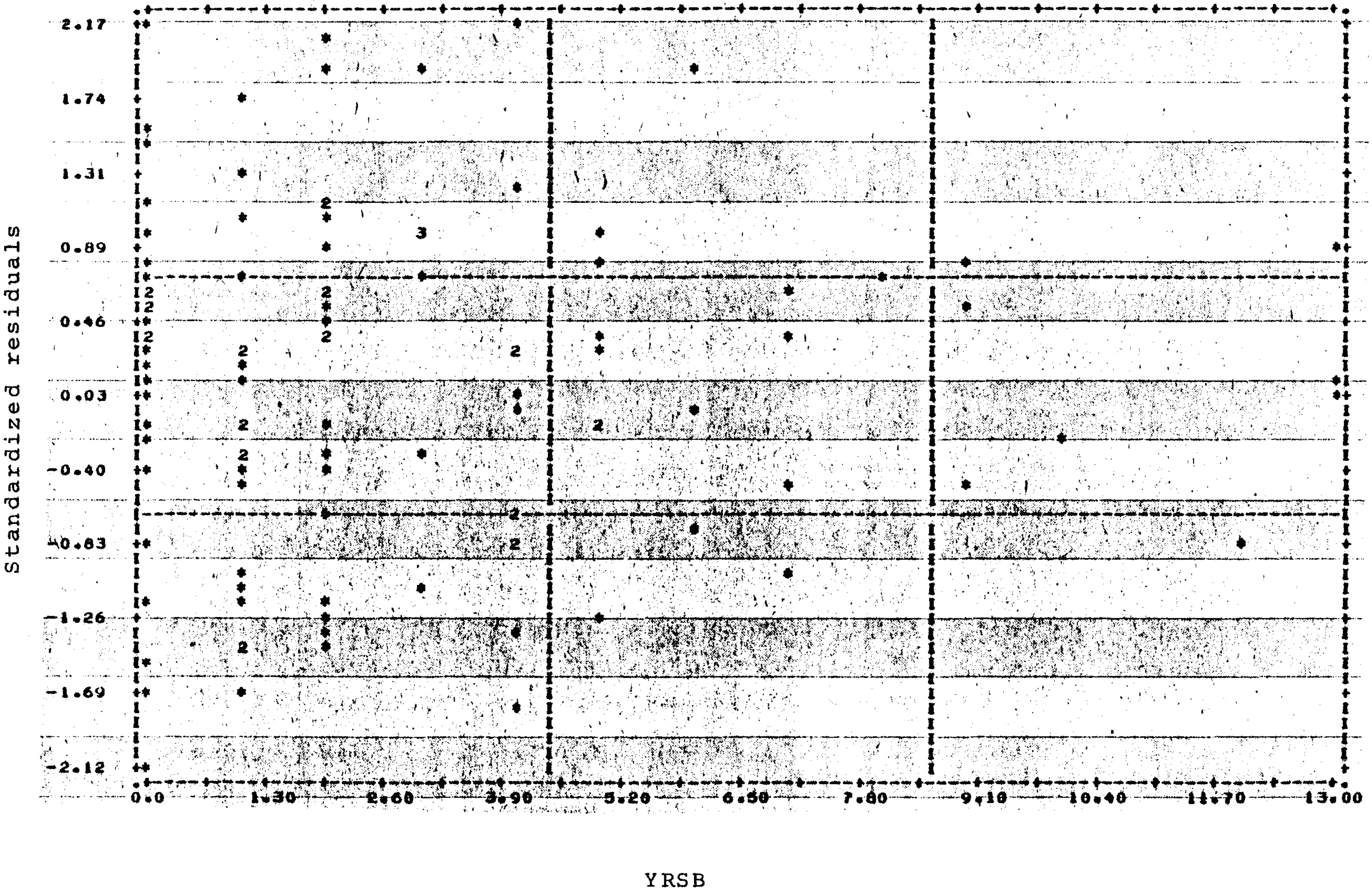


Figure 26. Graph of standardized residuals versus SCAV, regression with outliers removed, female sample.

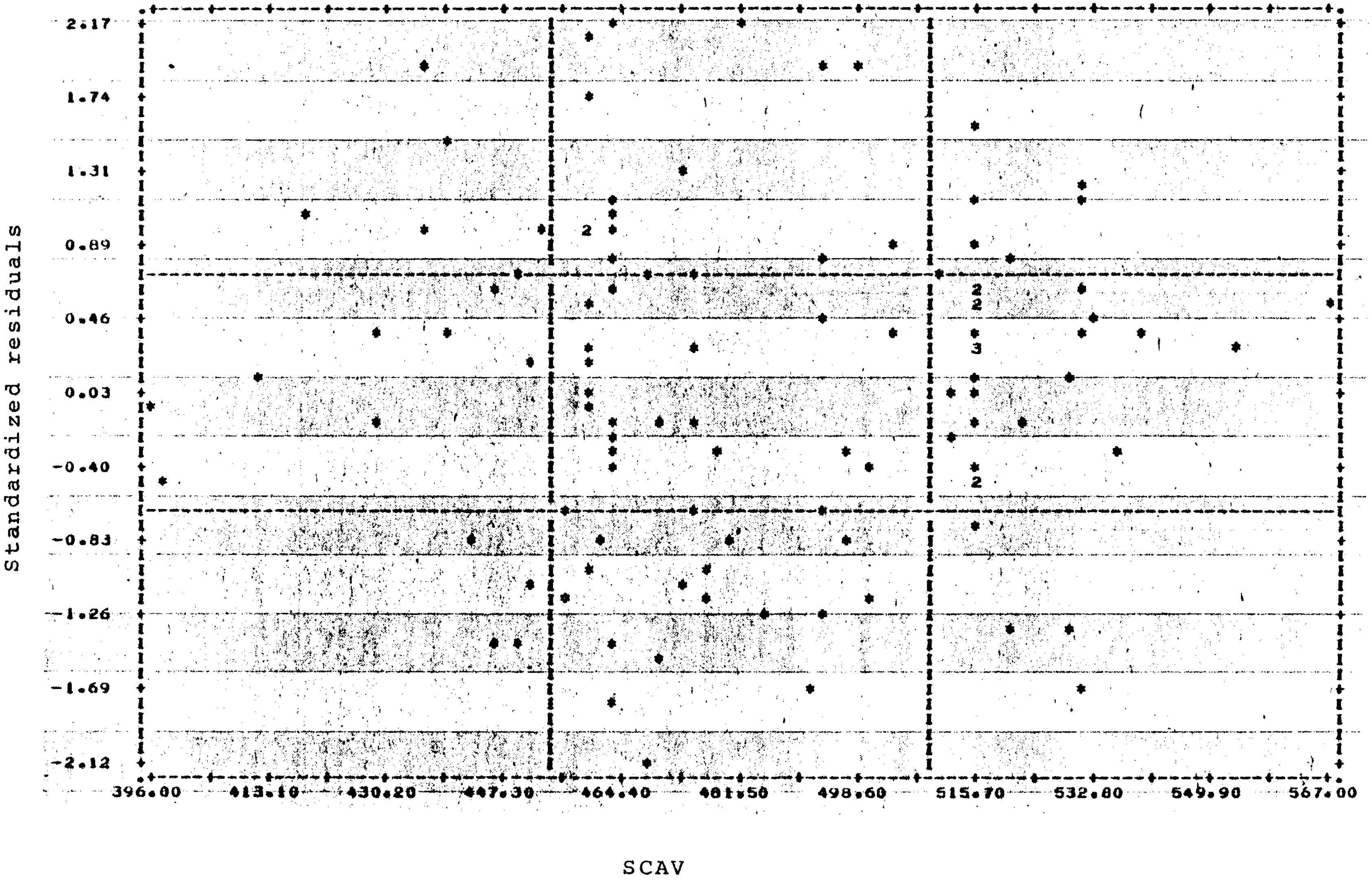


Figure 27. Graph of standardized residuals versus NOSC, regression with outliers removed, female sample.

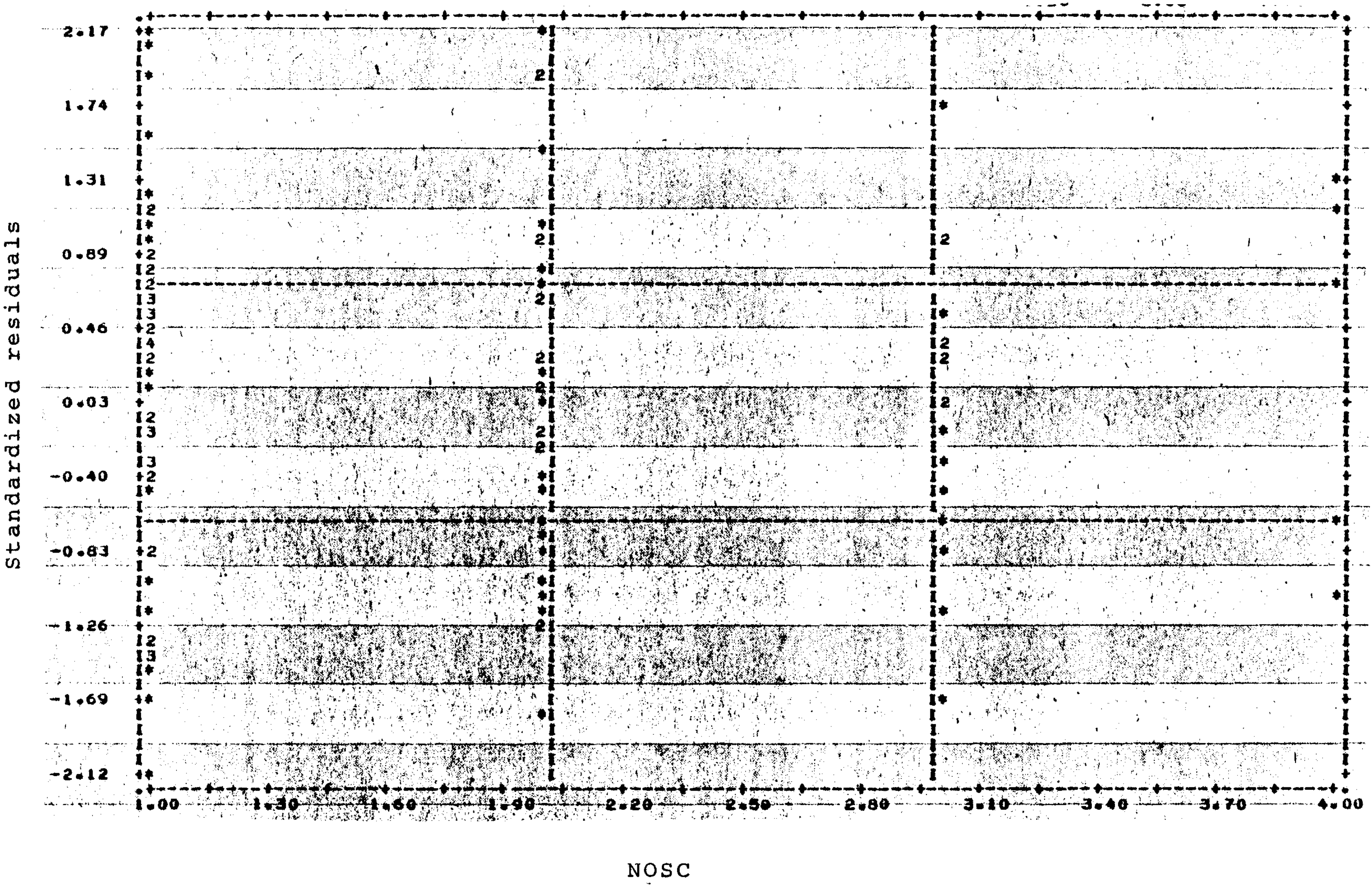


Figure 28. Graph of standardized residuals versus WORK, regression with outliers removed, female sample.

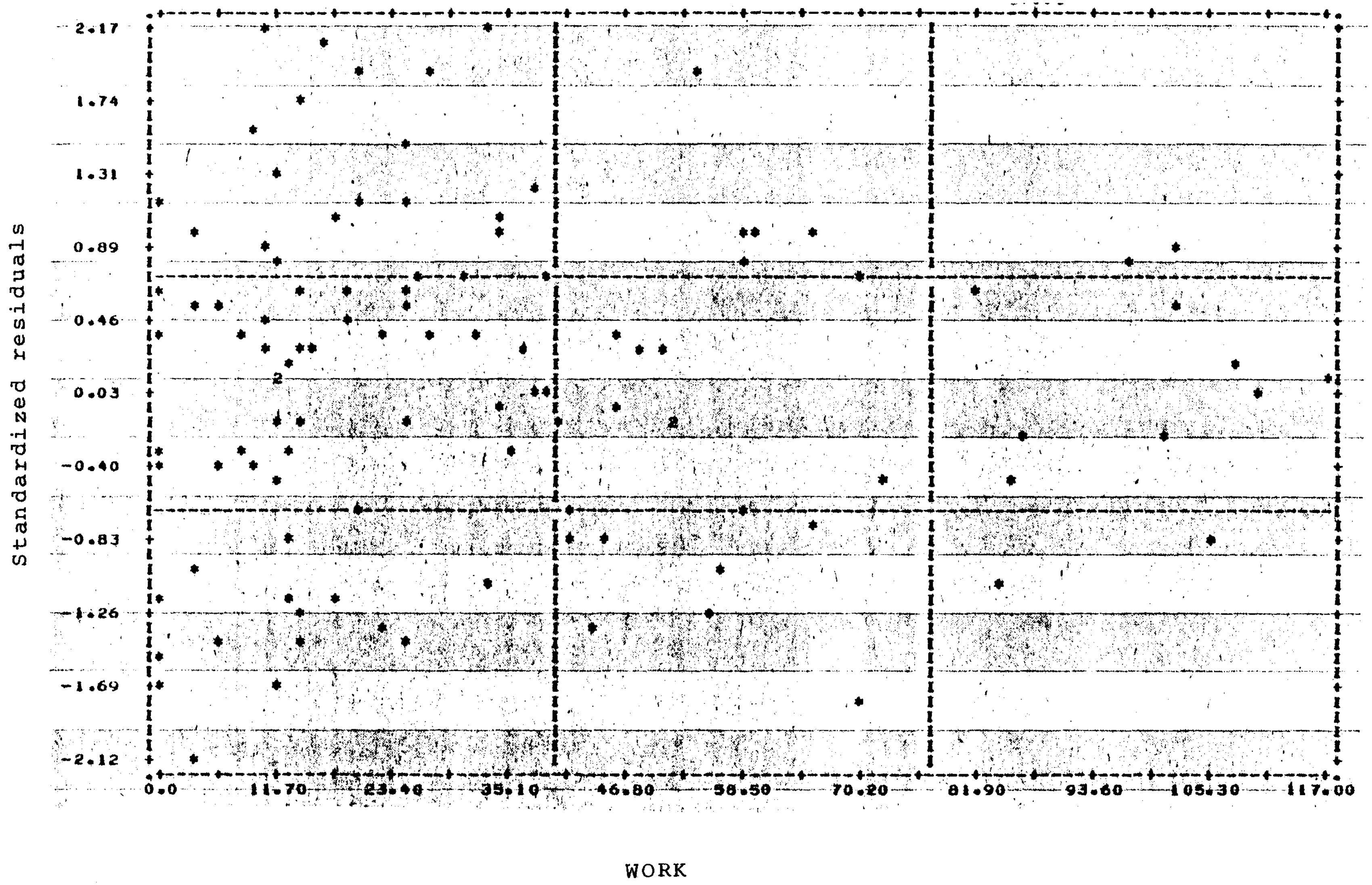


Figure 29. Graph of standardized residuals versus VERB, regression with outliers removed, female sample.

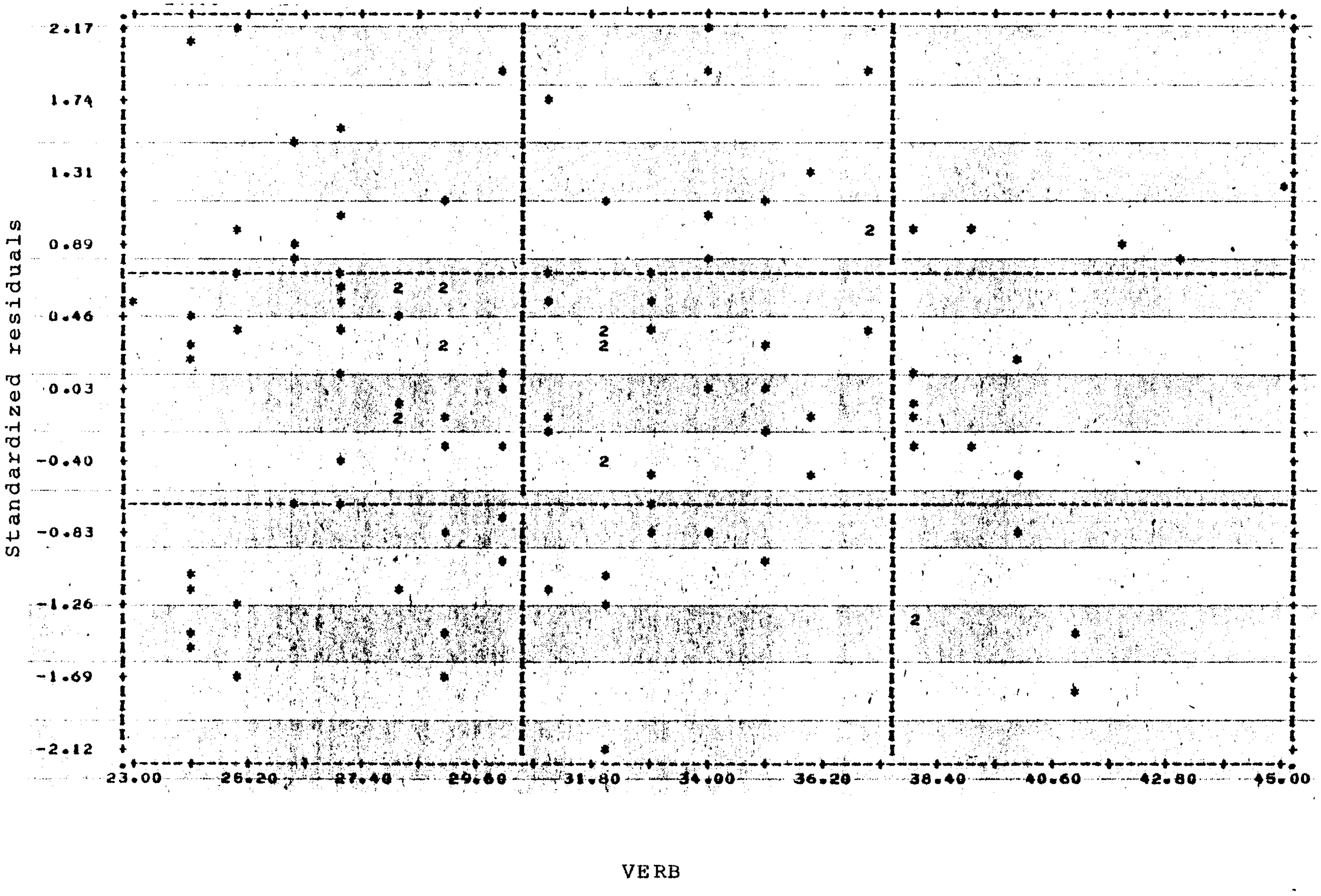
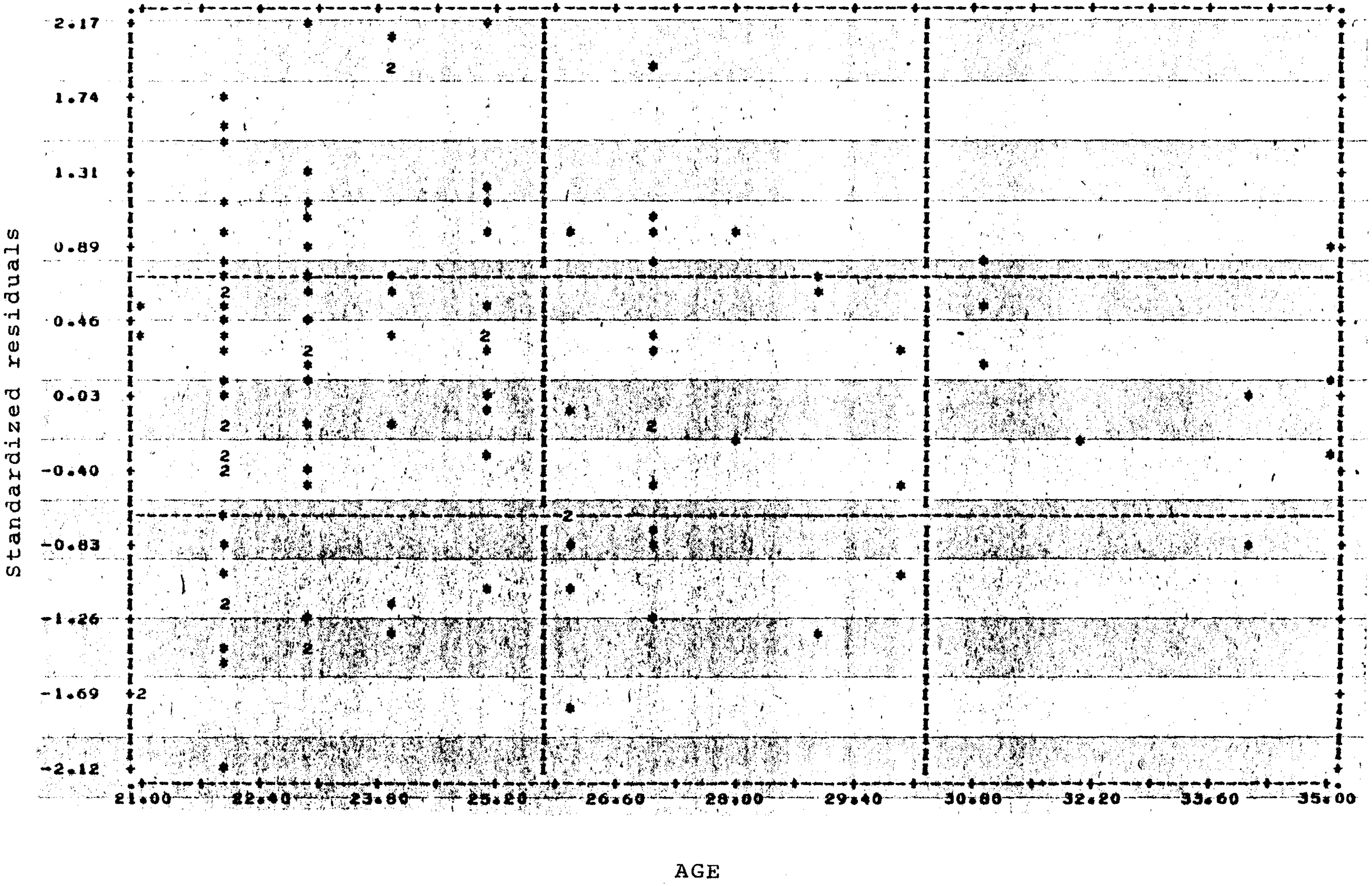


Figure 30. Graph of standardized residuals versus AGE, regression with outliers removed, female sample.



VERB, UGPA, NOSC and SCAV all show a random scatter, indicating a proper model fit for these variables. In the plots of WORK, AGE and YRSB, however, we see that the pattern of residuals seems to decrease in scatter as the independent variables increase, implying either the presence of heteroscedasticity in these three variables or that they have a curvilinear relationship with the dependent variable. A transformation or the addition of terms can be tried in the attempt to remove these model violations. Transformations for heteroscedasticity require the use of weighted least squares to solve for the parameter estimates. Before transforming the equations, the addition of other variables was tried to see if they could give any useful increase in the fit of the equation.

## Regressions with Added Terms for the Female Sample

The first additional variables added to the equation were two dummy variables, one indicating whether or not the student was an undergraduate business major, the second specifying whether the student held a degree beyond the bachelors, i.e., a masters or doctoral degree. These variables were designated as DUMM and OTDG, respectively. The results from this regression are shown in Table 11. The $R^2$ was .356, a slight improvement, and the standard error was .2648. OTDG entered second in the equation after QUAN, and accounted for 6% of the variation. DUMM entered

Table 11. Regression with Dummy Variables, Female Sample.

| VARIABLE | MULTIPLE R | R SQUARE | B | BETA |
|---|---|---|---|---|
| QUAN | 0.30340 | 0.09205 | 0.023465 | 0.35637 |
| OTDG | 0.39769 | 0.15816 | 0.175825 | 0.11697 |
| UGPA | 0.44928 | 0.20185 | 0.384526 | 0.39827 |
| YRSB | 0.53223 | 0.28327 | 0.009099 | 0.08907 |
| SCAV | 0.55230 | 0.30504 | 0.001703 | 0.18464 |
| NOSC | 0.57133 | 0.32642 | -0.069551 | -0.19509 |
| WORK | 0.59239 | 0.35093 | 0.003394 | 0.31780 |
| VERB | 0.59530 | 0.35438 | -0.004029 | -0.06582 |
| DUMM | 0.59677 | 0.35614 | -0.030814 | -0.04462 |
| (CONSTANT) |  |  | 0.192166 |  |

| | |
|---|---|
| MULTIPLE R | 0.59677 |
| R SQUARE | 0.35614 |
| STANDARD ERROR | 0.26477 |

| ANALYSIS OF VARIANCE | DF | SUM OF SQUARES | MEAN SQUARE | F |
|---|---|---|---|---|
| REGRESSION | 9 | 3.83890 | 0.42654 | 6.08440 |
| RESIDUAL | 99 | 6.94036 | 0.07010 |  |

the equation last and accounted for only .1%, one-tenth of one percent, of the variation, a relatively insignificant contribution.

The next group of variables added to the original equation were interaction terms. Five interaction terms were included. The first three of these include QUAN as one of the terms. Since the quantitative score on the GMAT has played an important part in the regression results, it is of interest to see whether its interaction with other variables will contribute to a major increase in the variance accounted for. Specifically, does the quantitative score interact sufficiently with the undergraduate grade point average, the school average, or the verbal score to influence $R^2$ to an extent worth entering the extra variables? Two additional interactions were investigated. These both contain AGE as one term. AGE was multiplied times work experience and years between to see if either of these combinations influences $R^2$ significantly beyond the individual entry of the specific terms. These interaction terms are QUAN*VERB, QUAN*SCAV, QUAN*UGPA, AGE*WORK and AGE*YRSB. Results of the regression are shown in Table 12. The value of $R^2$ was .376 and the standard error was .2647. All five interactions entered the equation. AGE*WORK, QUAN*SCAV and QUAN*VERB were given negative coefficients, so they act as suppressor variables to eliminate irrelevant variance in the other variables and thus improve the value

Table 12. Regression with Interaction Terms, Female Sample.

| VARIABLE | MULTIPLE R | R SQUARE | B | BETA |
|---|---|---|---|---|
| QUAN*UGPA | 0.40908 | 0.16735 | 0.037281 | 2.05842 |
| AGE*WORK | 0.50597 | 0.25600 | -0.000395 | -1.18457 |
| NOSC | 0.54064 | 0.29229 | -0.075410 | -0.21152 |
| SCAV | 0.56406 | 0.31817 | 0.003887 | 0.42148 |
| QUAN*SCAV | 0.59166 | 0.35006 | -0.000072 | -0.56608 |
| UGPA | 0.59670 | 0.35605 | -0.629579 | -0.65207 |
| WORK | 0.60223 | 0.36268 | 0.015151 | 1.41887 |
| AGE*YRSB | 0.60752 | 0.36908 | 0.002018 | 0.63768 |
| QUAN*VERB | 0.61106 | 0.37340 | -0.000289 | -0.19780 |
| QUAN | 0.61208 | 0.37464 | -0.047731 | -0.72492 |
| YRSB | 0.61284 | 0.37557 | -0.043503 | -0.42586 |
| VERB | 0.61290 | 0.37565 | 0.003561 | 0.05817 |
| (CONSTANT) | | | 1.992962 | |

| | |
|---|---|
| MULTIPLE R | 0.61290 |
| R SQUARE | 0.37565 |
| STANDARD ERROR | 0.26477 |

| ANALYSIS OF VARIANCE | DF | SUM OF SQUARES | MEAN SQUARE | F |
|---|---|---|---|---|
| REGRESSION | 12 | 4.04925 | 0.33744 | 4.81337 |
| RESIDUAL | 96 | 6.73001 | 0.07010 | |

of $R^2$. QUAN*UGPA contributed 16.7% of the variance of the dependent variable, AGE*WORK contributed 8%, QUAN*SCAV contributed 3%, AGE*YRSB contributed .6% and QUAN*VERB contributed .4%. QUAN*UGPA and AGE*WORK were the most valuable additions of the five interactions added.

When the deviation of the residuals pattern from a random scatter indicates some sort of curvature, several possibilities may occur. It may be that some important term is missing, causing, say, a systematic under- or over-prediction in a specific range which causes the residuals plot to appear to be curving down or up. A second possibility is that a transformation is necessary. A third alternative is that the relationship can be better expressed by a quadratic relationship which is showing up as a parabolic scatterplot. This last case is dealt with by the addition of higher order variable terms to the equation. If a squared term is added to the equation, and is necessary, then this should result in a better fit and an improved pattern in the residuals plot.

Several squared terms were added to see their effect on the regression equation. As with the interaction terms, the increase in $R^2$ should be significant enougn to justify the inclusion of additional terms. The terms added were AGE*AGE, WORK*WORK and YRSB*YRSB. The regression results are shown in Table 13. The final $R^2$ was .379 with a standard error of .2626. YRSB*YRSB added 10.7% to the

Table 13. Regression with Three Squared Terms, Female Sample.

| VARIABLE | MULTIPLE R | R SQUARE | B | BETA |
|---|---|---|---|---|
| QUAN | 0.30340 | 0.09205 | 0.023248 | 0.35307 |
| UGPA | 0.39558 | 0.15648 | 0.424669 | 0.43984 |
| YRSB*YRSB | 0.51387 | 0.26407 | 0.010379 | 1.13636 |
| NOSC | 0.53262 | 0.28369 | -0.066662 | -0.18698 |
| WORK | 0.55370 | 0.30659 | 0.008118 | 0.76024 |
| SCAV | 0.58172 | 0.33839 | 0.002036 | 0.22078 |
| WORK*WORK | 0.59414 | 0.35300 | -0.000072 | -0.70758 |
| YRSB | 0.60014 | 0.36017 | -0.087492 | -0.85648 |
| VERB | 0.60225 | 0.36270 | -0.003813 | -0.06228 |
| AGE | 0.60277 | 0.36334 | 0.374349 | 3.99592 |
| AGE*AGE | 0.61598 | 0.37943 | -0.006666 | -3.86497 |
| (CONSTANT) | | | -5.135230 | |

| | |
|---|---|
| MULTIPLE R | 0.61598 |
| R SQUARE | 0.37943 |
| STANDARD ERROR | 0.26261 |

| ANALYSIS OF VARIANCE | DF | SUM OF SQUARES | MEAN SQUARE | F |
|---|---|---|---|---|
| REGRESSION | 11 | 4.08999 | 0.37182 | 5.39167 |
| RESIDUAL | 97 | 6.68926 | 0.06898 | |

percent of variance accounted for, WORK*WORK added 1.4% and AGE*AGE added 1.6%. WORK*WORK and AGE*AGE had negative parameter values.

The addition of variable terms has not increased the value of $R^2$ to any great extent. The original $R^2$ (without outliers) was .341. The values of $R^2$ with additional terms were .356, .357 and .379. Several terms, however, added more than .05 to the value of $R^2$. These were OTDG, YRSB*YRSB, QUAN*UGPA and AGE*WORK.

A fourth regression with additional terms was run using the original eight variables plus OTDG, YRSB*YRSB, QUAN*UGPA and AGE*WORK. The results of the run are shown in Table 14. $R^2$ had a value of .389 and the standard error was .2619. $R^2$ was improved from 13% to almost 39% through the elimination of outliers and the addition of other variable terms. However, the addition of terms has not been able to correct the model violations indicated in the residuals plots. The residuals were standardized and plotted against each variable. The results, shown in Figures 31 through 43, indicate model violations remaining in those terms involving AGE, YRSB and WORK. Specifically, Figures 33, 40, 42 and 43 indicate model violations remaining in the terms AGE*WORK, YRSB*YRSB, YRSB and AGE. These show a decrease in residual variability as the independent variable increases. The next step was to attempt the removal of the violations with a transformation.

Table 14. Regression with Interactions,Dummy Variable and Squared Term, Female Sample.

| VARIABLE | MULTIPLE R | R SQUARE | B | BETA |
|---|---|---|---|---|
| QUAN*UGPA | 0.40908 | 0.16735 | 0.038198 | 2.10907 |
| AGE*WORK | 0.50597 | 0.25600 | -0.000636 | -1.90280 |
| NOSC | 0.54064 | 0.29229 | -0.073374 | -0.20581 |
| SCAV | 0.56406 | 0.31817 | 0.001833 | 0.19871 |
| QUAN | 0.59033 | 0.34849 | -0.095068 | -1.44385 |
| OTDG | 0.60037 | 0.36044 | 0.166471 | 0.11075 |
| UGPA | 0.60759 | 0.36917 | -0.703491 | -0.72862 |
| WORK | 0.61016 | 0.37229 | 0.021186 | 1.98398 |
| YRSB*YRSB | 0.61634 | 0.37987 | 0.005264 | 0.57631 |
| VERB | 0.62025 | 0.38471 | -0.004328 | -0.07069 |
| YRSB | 0.62233 | 0.38730 | -0.033312 | -0.32610 |
| AGE | 0.62382 | 0.38915 | 0.011042 | 0.11787 |
| (CONSTANT) | | | 3.252225 | |

| | |
|---|---|
| MULTIPLE R | 0.62382 |
| R SQUARE | 0.38915 |
| STANDARD ERROR | 0.26189 |

| ANALYSIS OF VARIANCE | DF | SUM OF SQUARES | MEAN SQUARE | F |
|---|---|---|---|---|
| REGRESSION | 12 | 4.19475 | 0.34956 | 5.09651 |
| RESIDUAL | 96 | 6.58450 | 0.06859 | |

Figure 31. Graph of standardized residuals versus $\hat{y}$, regression with added terms, female sample.

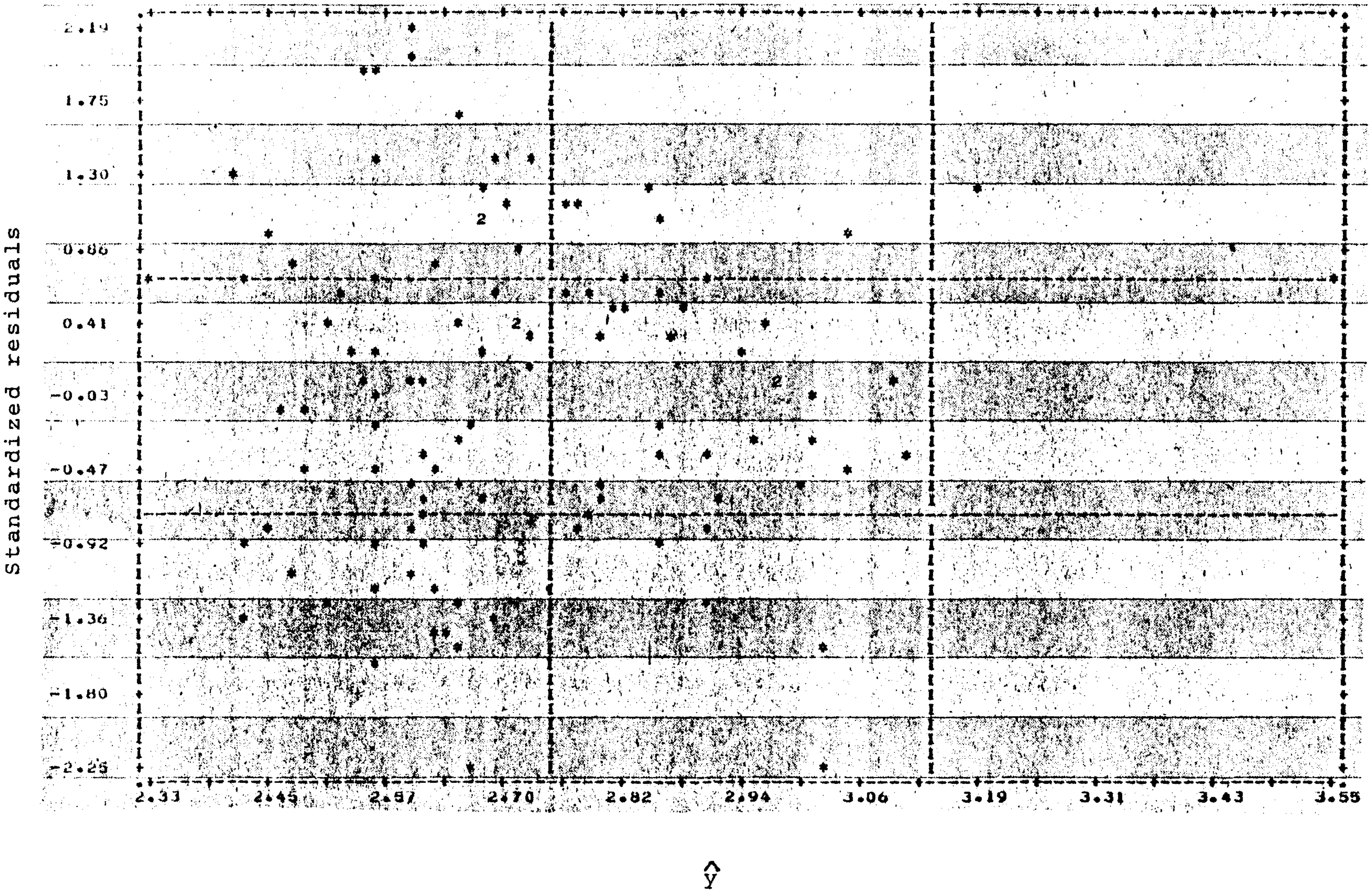


Figure 32. Graph of standardized residuals versus QUAN*UGPA, regression with added terms, female sample.

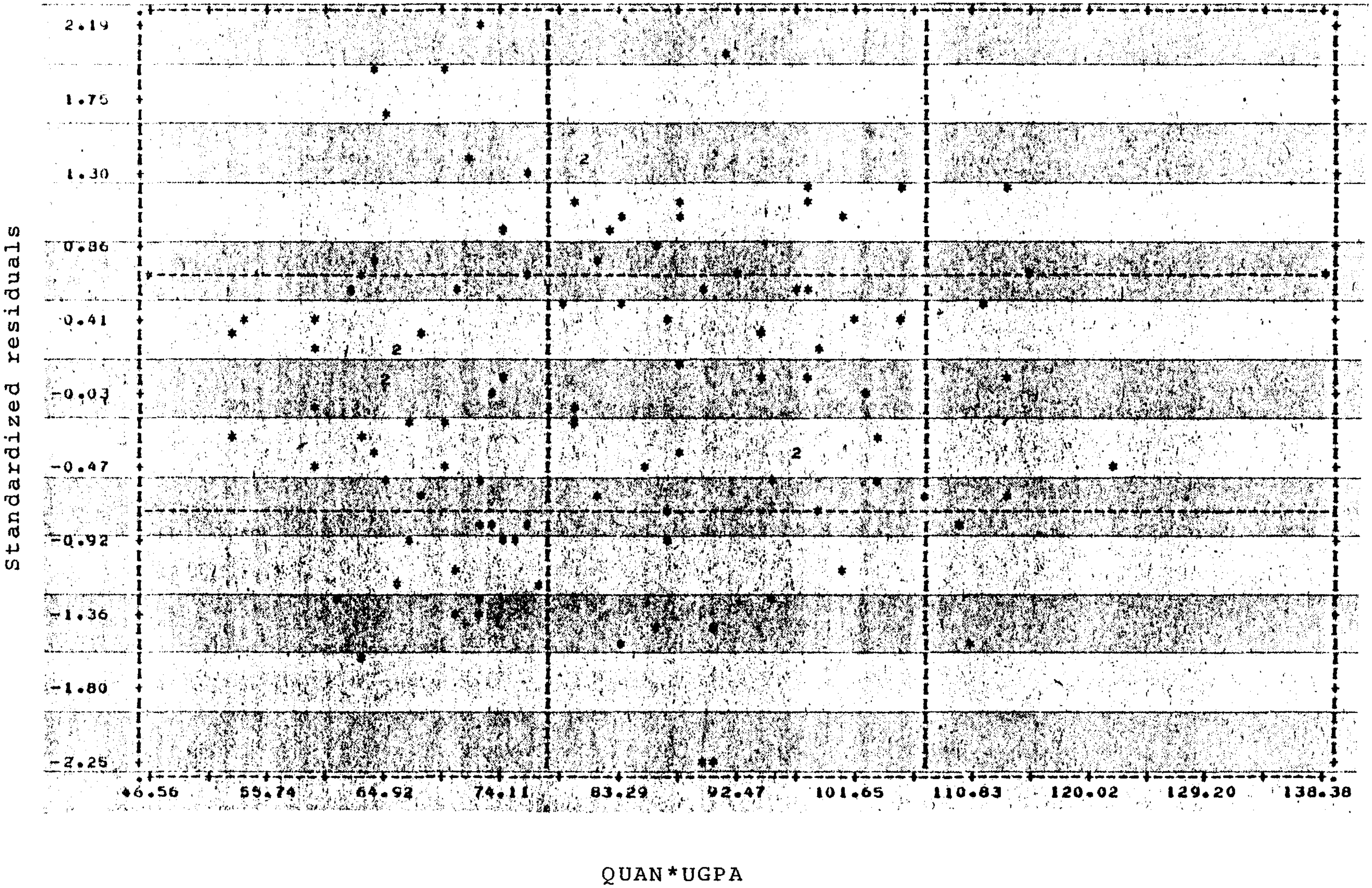


Figure 33. Graph of standardized residuals versus AGE*WORK, regression with added terms, female sample.

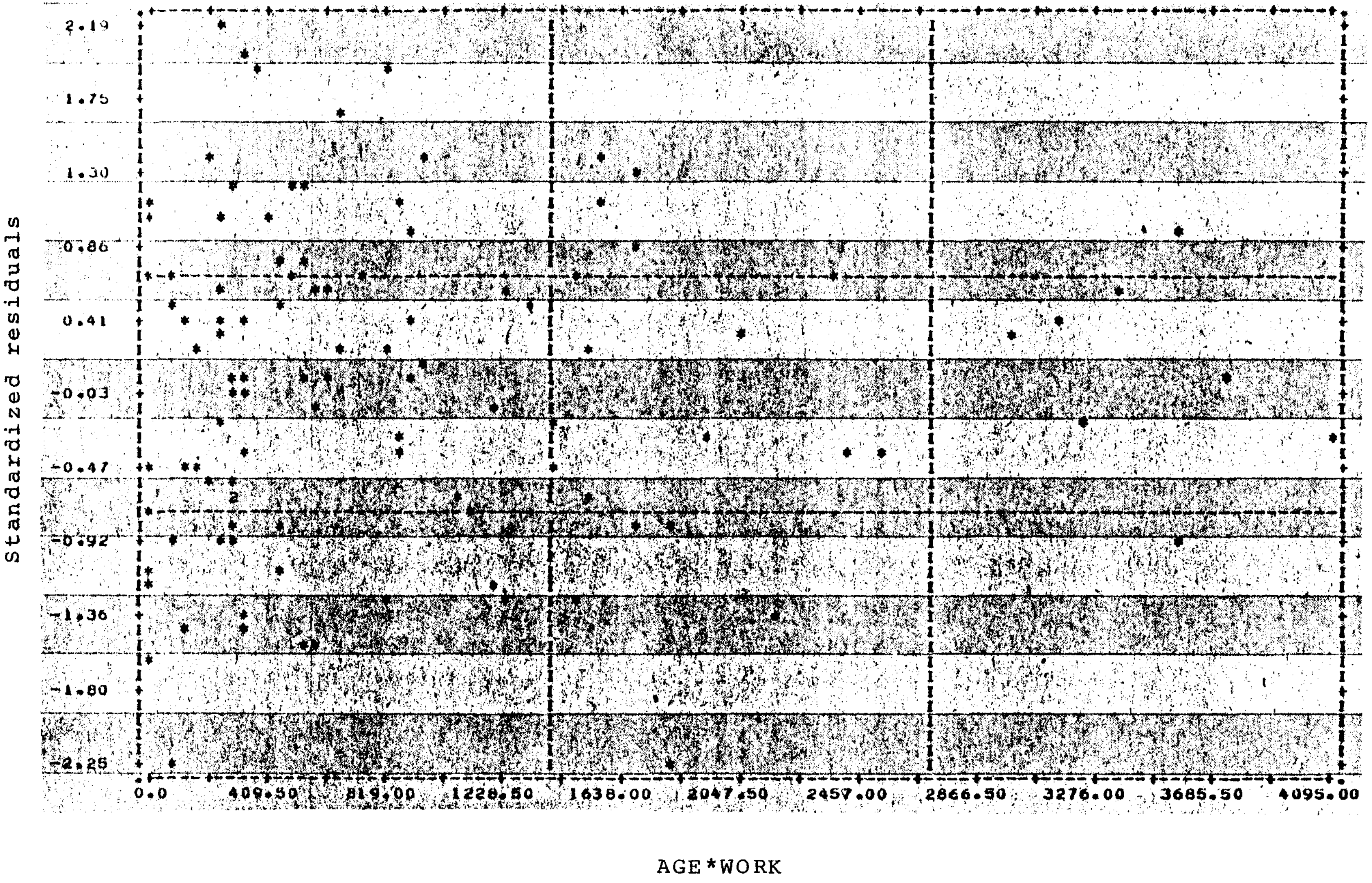


Figure 34. Graph of standardized residuals versus NOSC, regression with added terms, female sample.

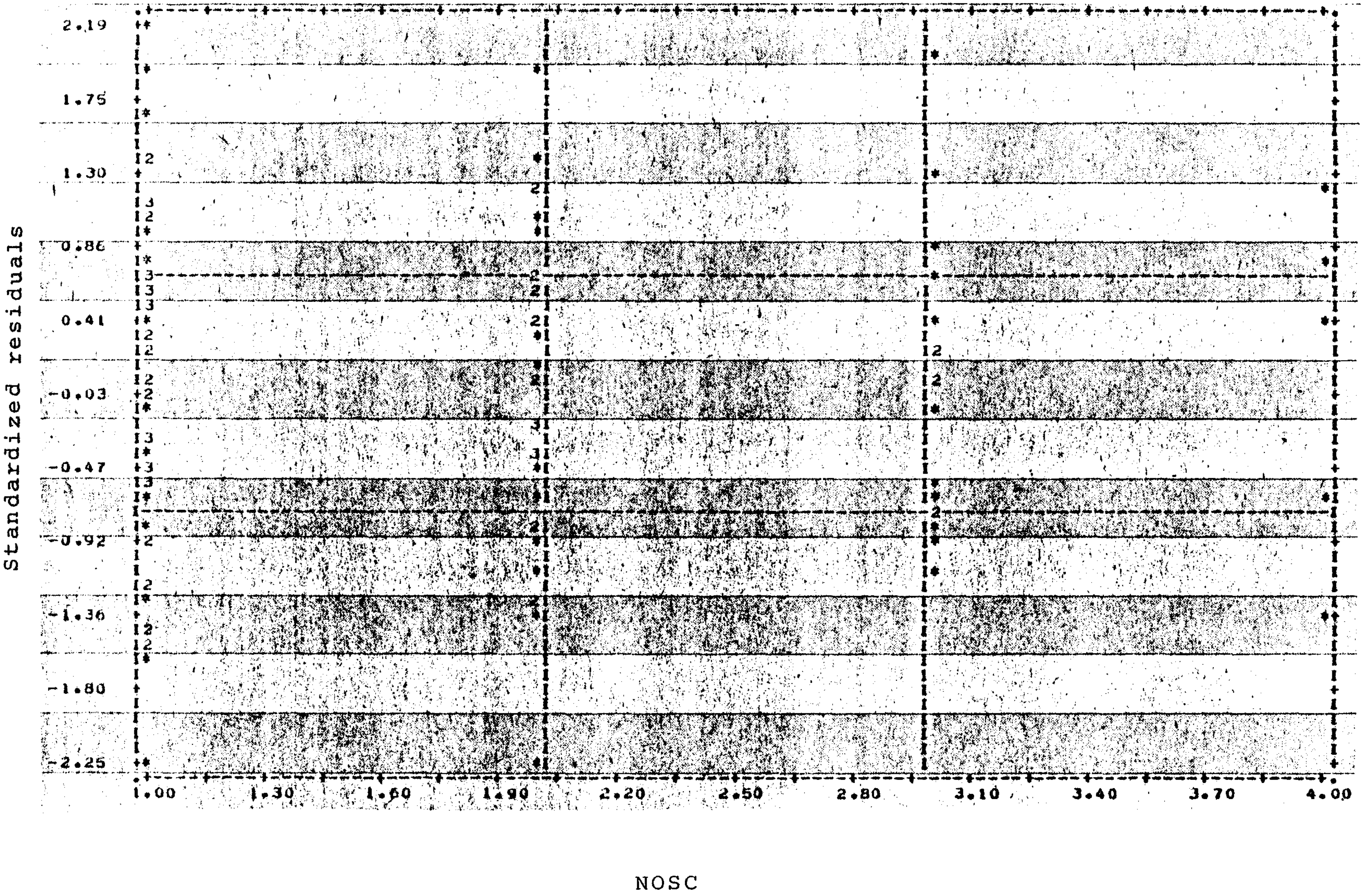


Figure 35. Graph of standardized residuals versus SCAV, regression with added terms, female sample.

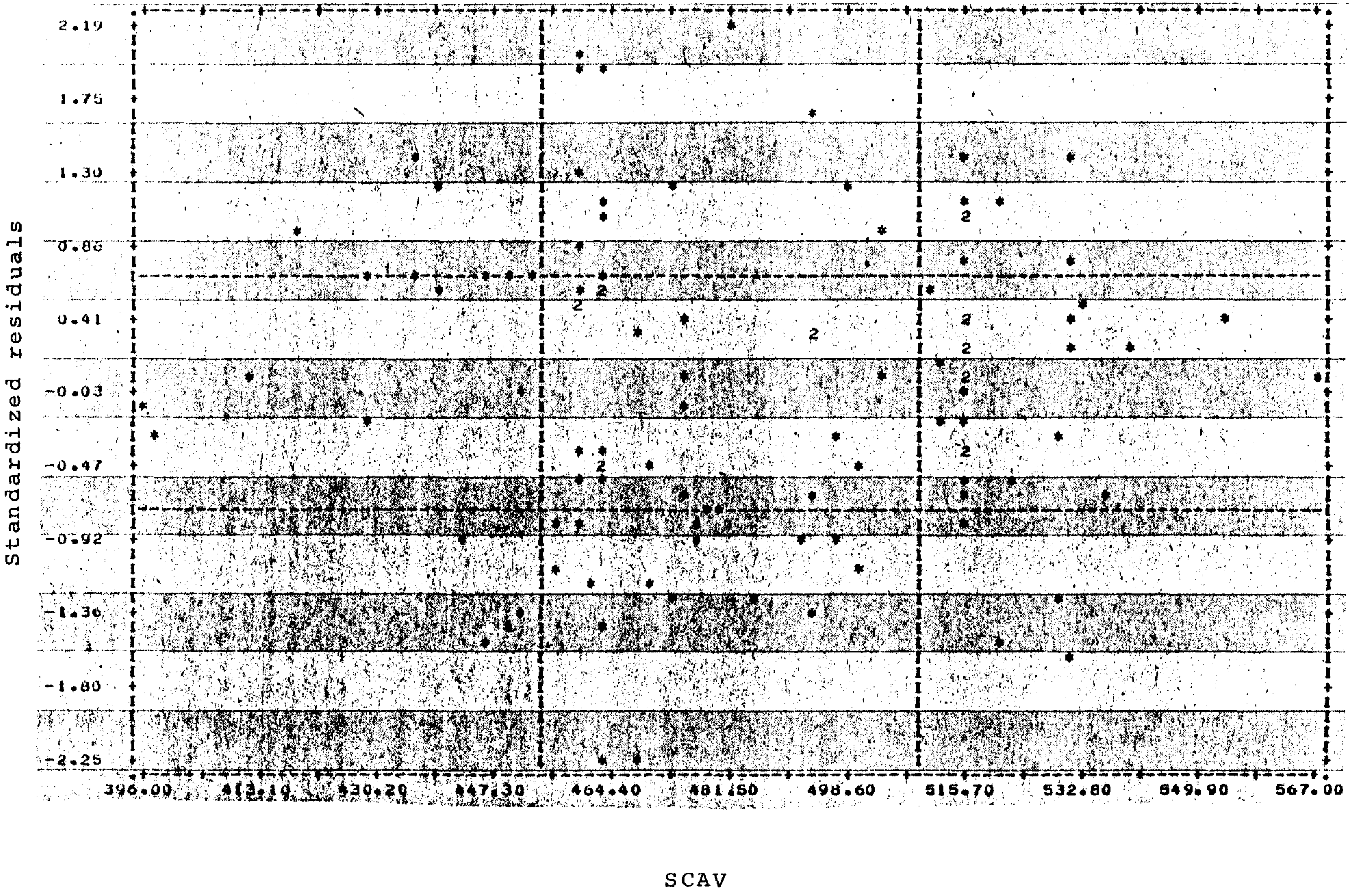


Figure 36. Graph of standardized residuals versus QUAN, regression with added terms, female sample.

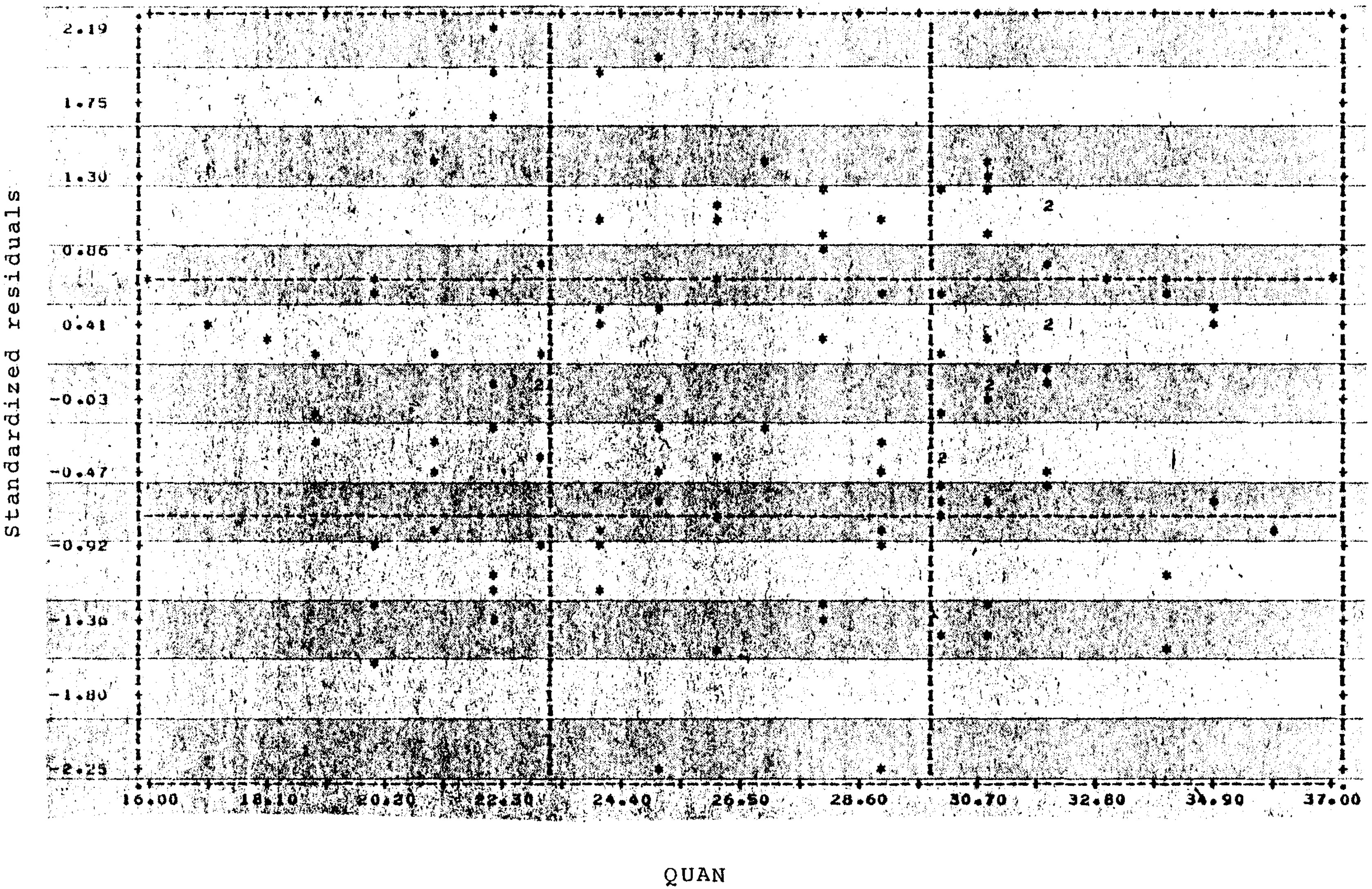


Figure 37. Graph of standardized residuals versus OTDG, regression with added terms, female sample.

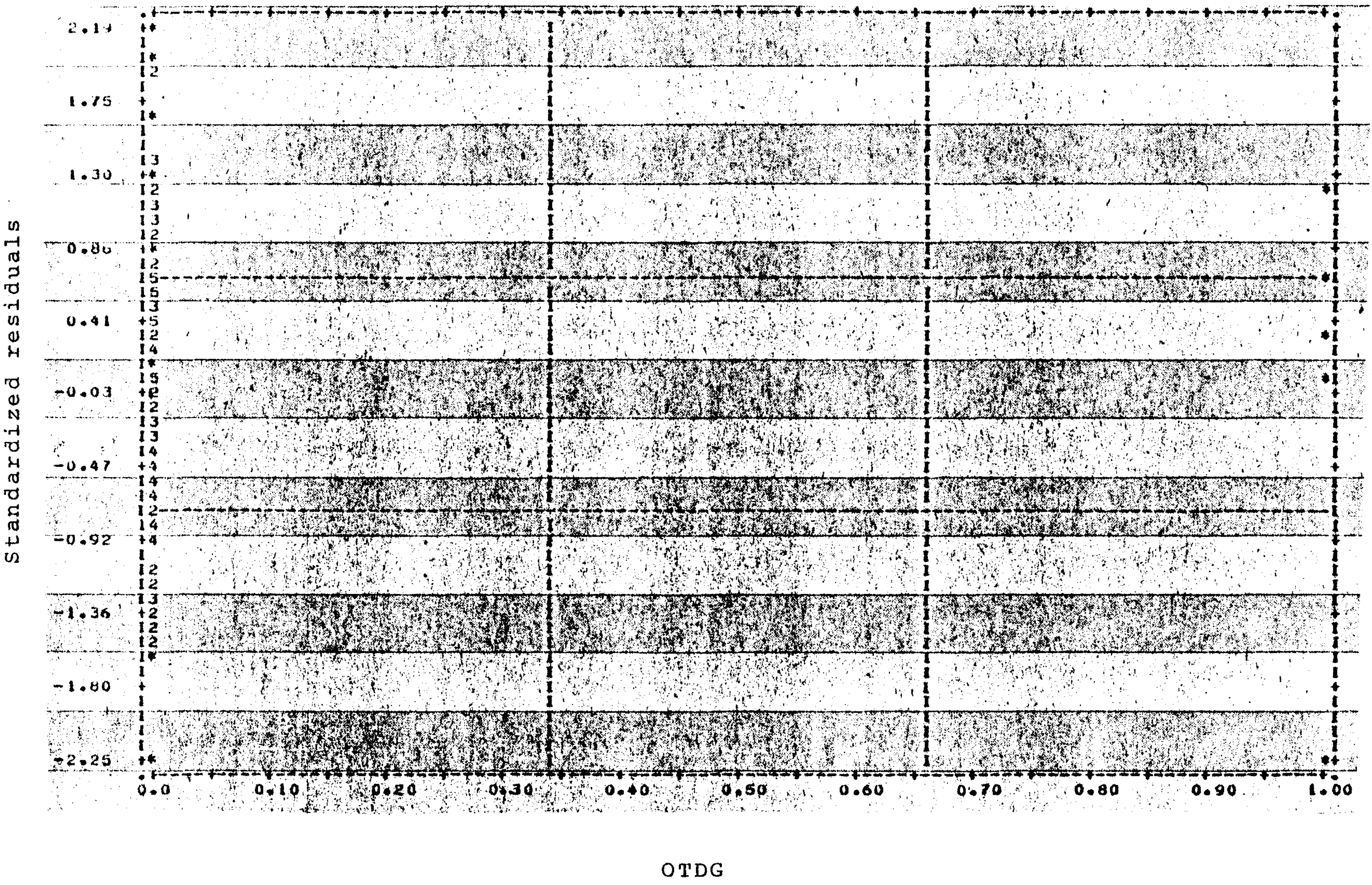


Figure 38. Graph of standardized residuals versus UGPA, regression with added terms, female sample.

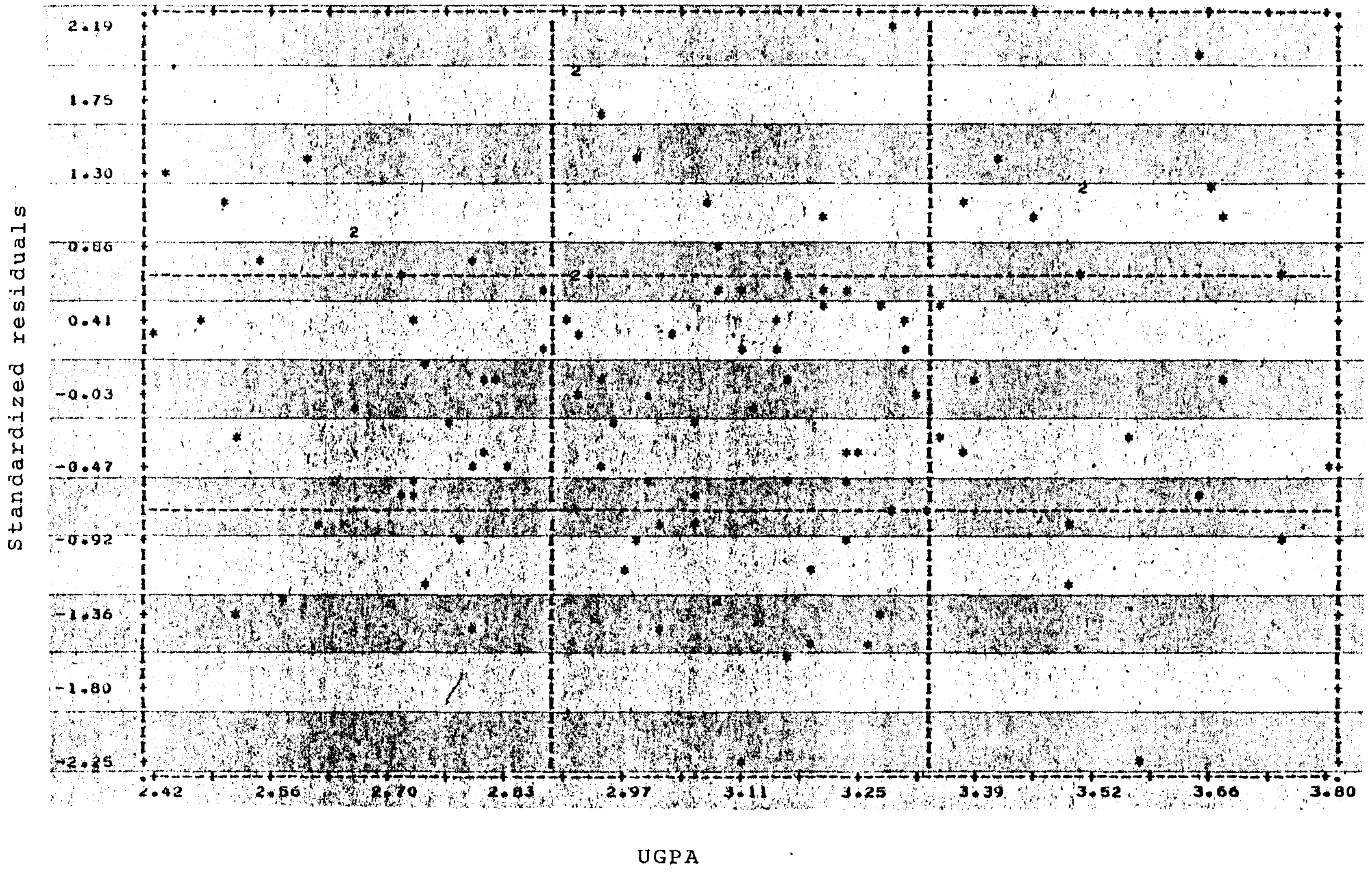


Figure 39. Graph of standardized residuals versus WORK, regression with added terms, female sample.

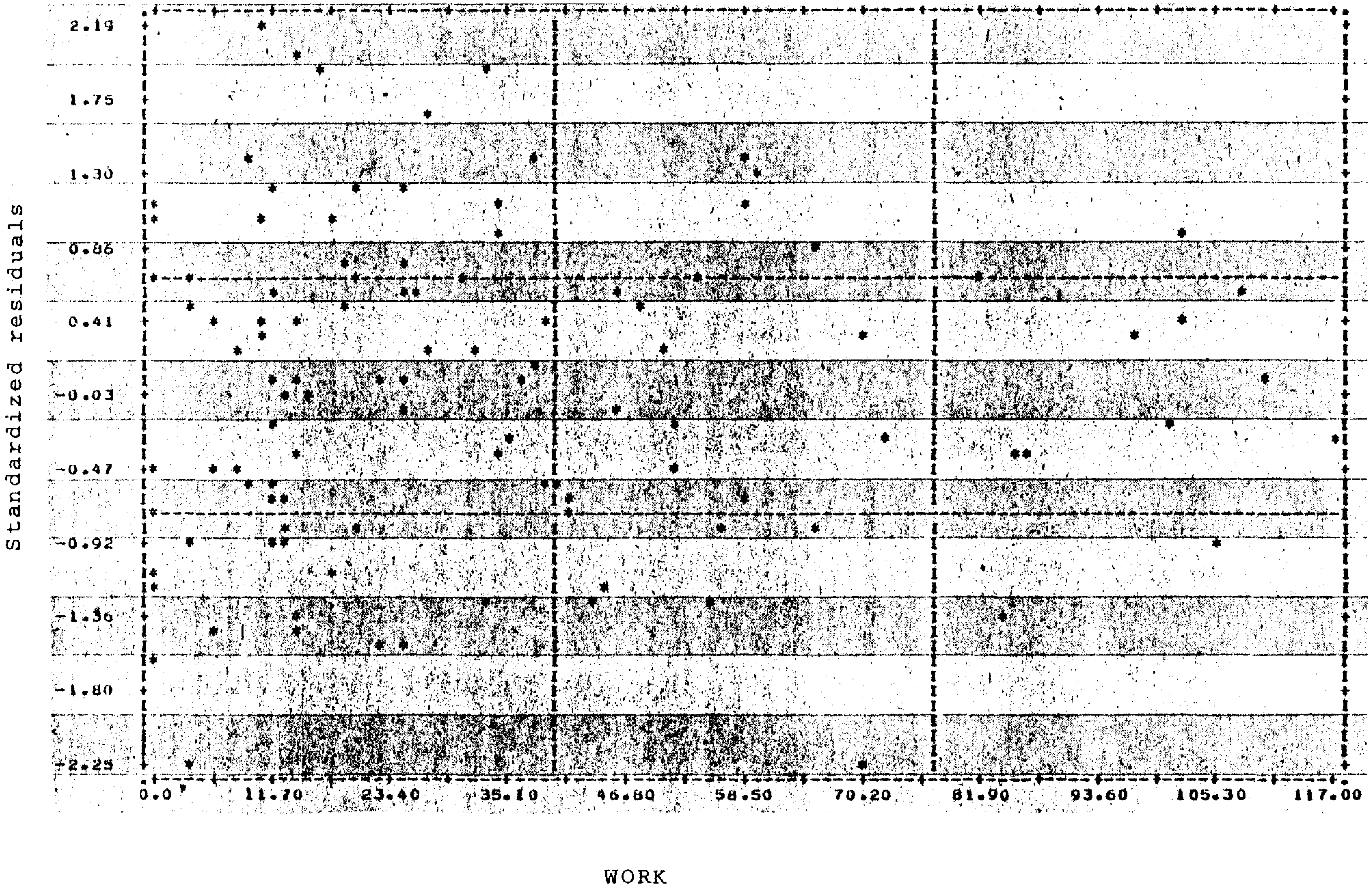


Figure 40. Graph of standardized residuals versus YRSB*YRSB, regression with added terms, female sample.

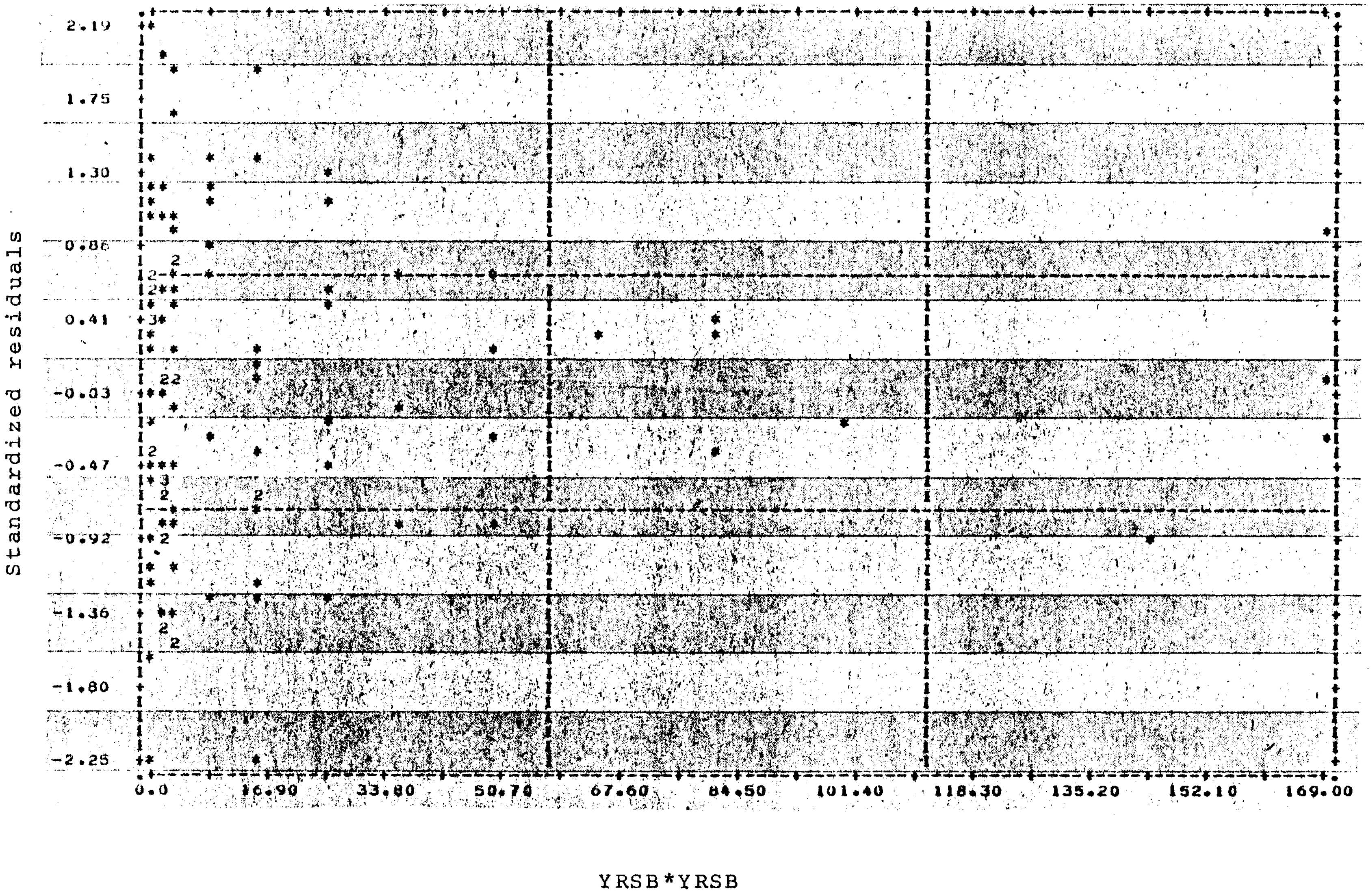


Figure 41. Graph of standardized residuals versus VERB, regression with added terms, female sample.

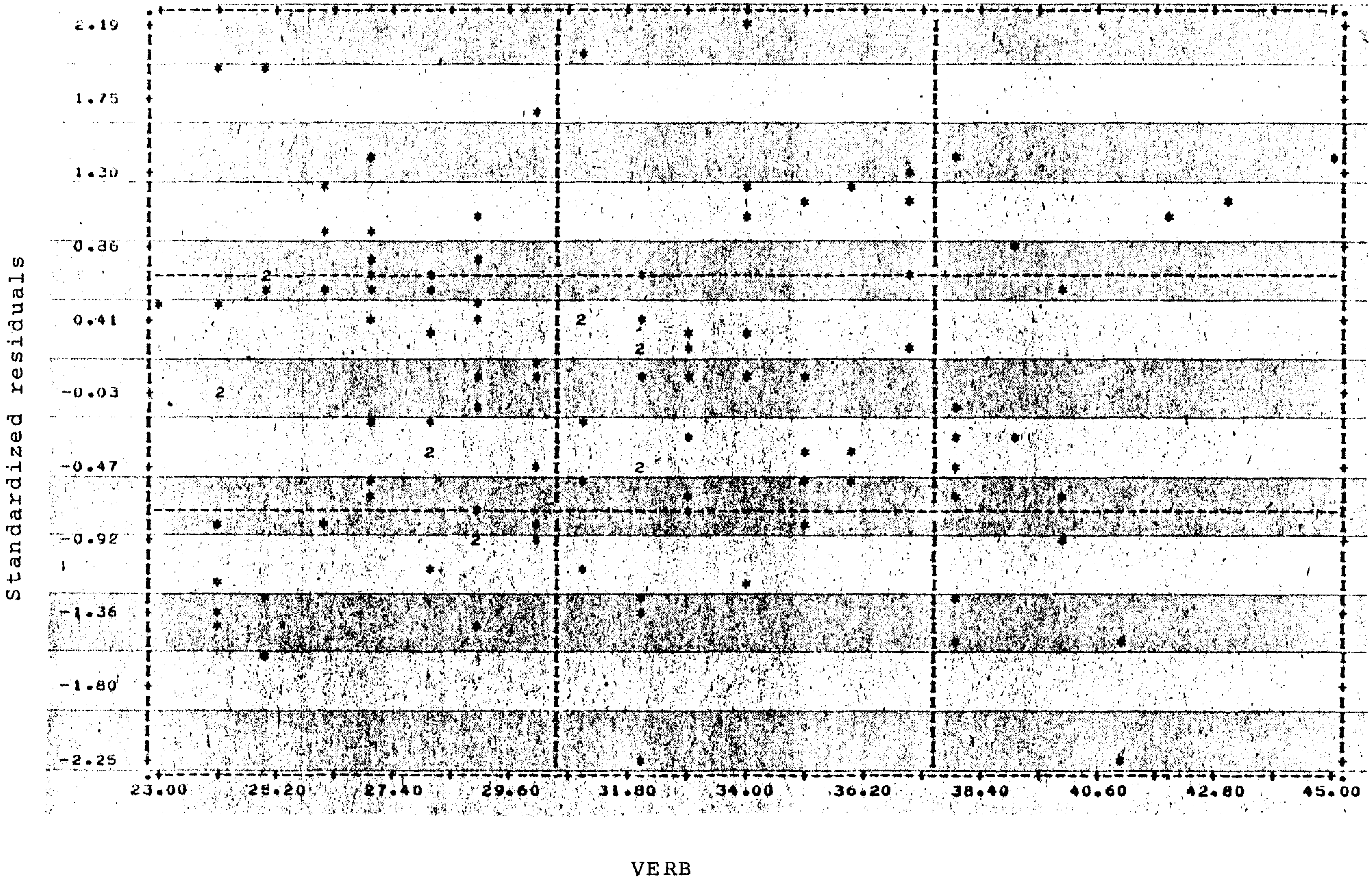


Figure 42. Graph of standardized residuals versus YRSB, regression with added terms, female sample.

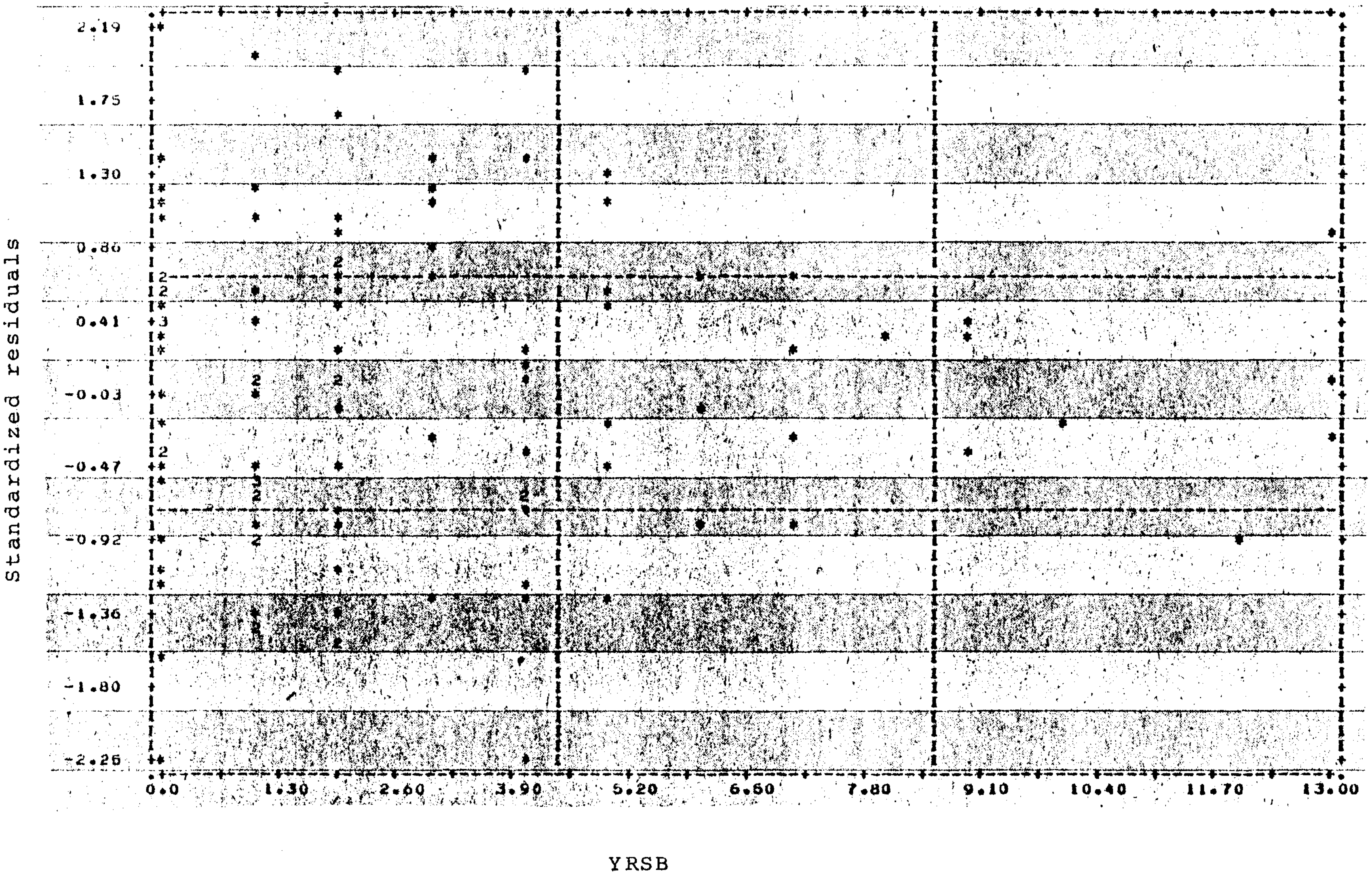
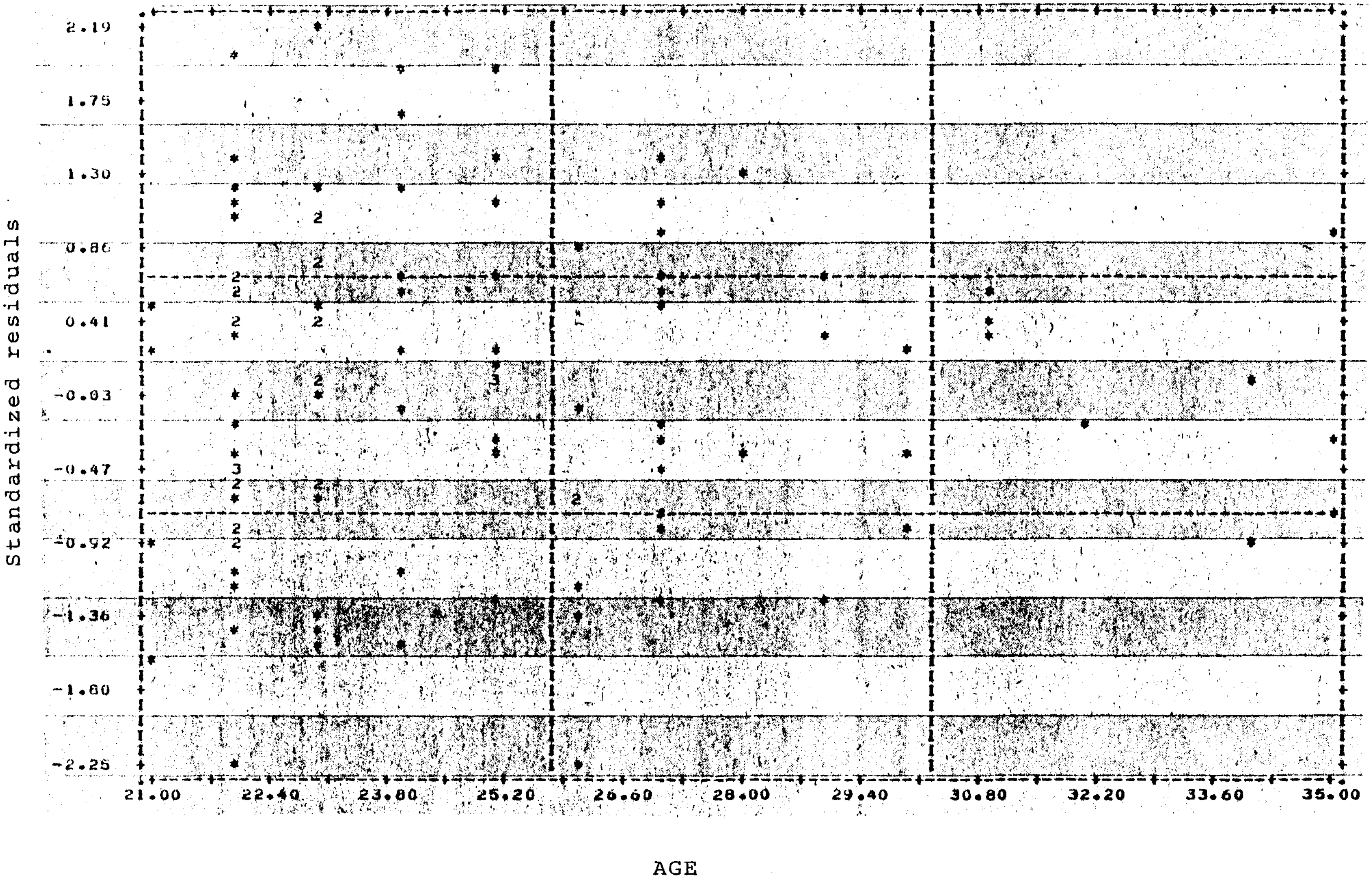


Figure 43. Graph of standardized residuals versus AGE, regression with added terms, female sample.



## Regression with Transformed Variables and Residuals Plots for the Female Sample

As previously mentioned, AGE, YRSB and WORK all have residuals plots which indicate some model violation or misspecification. Since added terms have not corrected these violations, it may be that this heteroscedasticity can be removed through the application of a transformation. This is the method of weighted least squares. Since AGE, YRSB and WORK are interrelated concepts, it is possible that one transformation may correct the heteroscedasticity in all three variables. The transformation to be used is the following: all terms in the original regression equation will be multiplied by 10/(V - AGE) where V is the value AGE approaches as the residuals plot decreases in scatter. In this case, V has a value of 36. The transformed equation is:

$$\begin{aligned} 10*GGPA/(36-AGE) = {} & 10*\beta_0/(36-AGE) + \beta_1 10*UGPA/(36-AGE) \\ & + \beta_2 10*YRSB/(36-AGE) + \beta_3 10*NOSC/(36-AGE) \\ & + \beta_4 10*WORK/(36-AGE) + \beta_5 10*AGE/(36-AGE) \\ & + \beta_6 10*SCAV/(36-AGE) + \beta_7 10*VERB/(36-AGE) \\ & + \beta_8 10*QUAN/(36-AGE). \end{aligned}$$

The regression program was run on the transformed equation. The results from this run are shown in Table 15. $R^2$ had a value of .9967. The fitted equation generated by this program is:

Table 15. Regression with Transformation, Female Sample.

| VARIABLE | MULTIPLE R | R SQUARE | B | BETA |
|---|---|---|---|---|
| 10*SCAV/(36-AGE) | 0.99103 | 0.98214 | 0.003750 | 0.64728 |
| 10*YRSB/(36-AGE) | 0.99435 | 0.98873 | 0.004686 | 0.01926 |
| 10*QUAN/(36-AGE) | 0.99760 | 0.99520 | 0.035569 | 0.32652 |
| 10*NOSC/(36-AGE) | 0.99790 | 0.99581 | -0.112331 | -0.08063 |
| 10*UGPA/(36-AGE) | 0.99804 | 0.99608 | 0.323445 | 0.31443 |
| 10*WORK/(36-AGE) | 0.00817 | 0.99635 | 0.004611 | 0.16089 |
| 10*VERB/(36-AGE) | 0.99832 | 0.99665 | -0.012998 | -0.14163 |
| 10/(36-AGE) | 0.99836 | 0.99672 | -0.674741 | -0.23142 |
| (CONSTANT) | | | 0.061871 | |

| | |
|---|---|
| MULTIPLE R | 0.99836 |
| R SQUARE | 0.99672 |
| STANDARD ERROR | 0.28149 |

| ANALYSIS OF VARIANCE | DF | SUM OF SQUARES | MEAN SQUARE | F |
|---|---|---|---|---|
| REGRESSION | 8 | 2362.64932 | 295.33116 | 3727.22081 |
| RESIDUAL | 100 | 7.76516 | 0.07924 | |

$$(3)\quad 10\widehat{GGPA}/(36\text{-}AGE) = .0375SCAV/(36\text{-}AGE) + .0459YRSB/(36\text{-}AGE) + .3557QUAN/(36\text{-}AGE) + 1.123NOSC/(36\text{-}AGE) + 3.234UGPA/(36\text{-}AGE) + .0461WORK/(36\text{-}AGE) - .130VERB/(36\text{-}AGE) - 6.747/(36\text{-}AGE) + .0619$$

with a standard error of .2815. To recover the value of GGPA, we multiply the predicted value by (36-AGE)/10. For example, given the student with the following variable values

| | |
|---|---|
| SCAV | 480 |
| YRSB | 3 |
| QUAN | 27 |
| NOSC | 2 |
| UGPA | 3.0 |
| WORK | 38 |
| VERB | 31 |

this equation would predict

$$10\widehat{GGPA}/(36\text{-}AGE) = 1.64 + .01 + .87 - .20 + .88 + .16 - .37 - .56 + .0619 = 2.49$$

So, $\widehat{GGPA} = (10\widehat{GGPA}/(36\text{-}AGE))((36\text{-}AGE)/10) = (2.49)(1.1) = 2.74$. This example was calculated for a student of age 25.

The residuals plots of the regression using a transformed equation are shown in Figures 44 through 52. These graphs show the majority of points falling in the first third of each graph. A few points, mainly from individuals who are older or with more work experience fall in the right-hand third or two-thirds. The plots are randomly distributed with all but two points falling between plus and minus two. In the graphs where only a few points fall in the right-hand two-thirds, $\hat{y}$, SCAV,

Figure 44. Graph of standardized residuals versus $\hat{y}$, transformed regression, female sample.

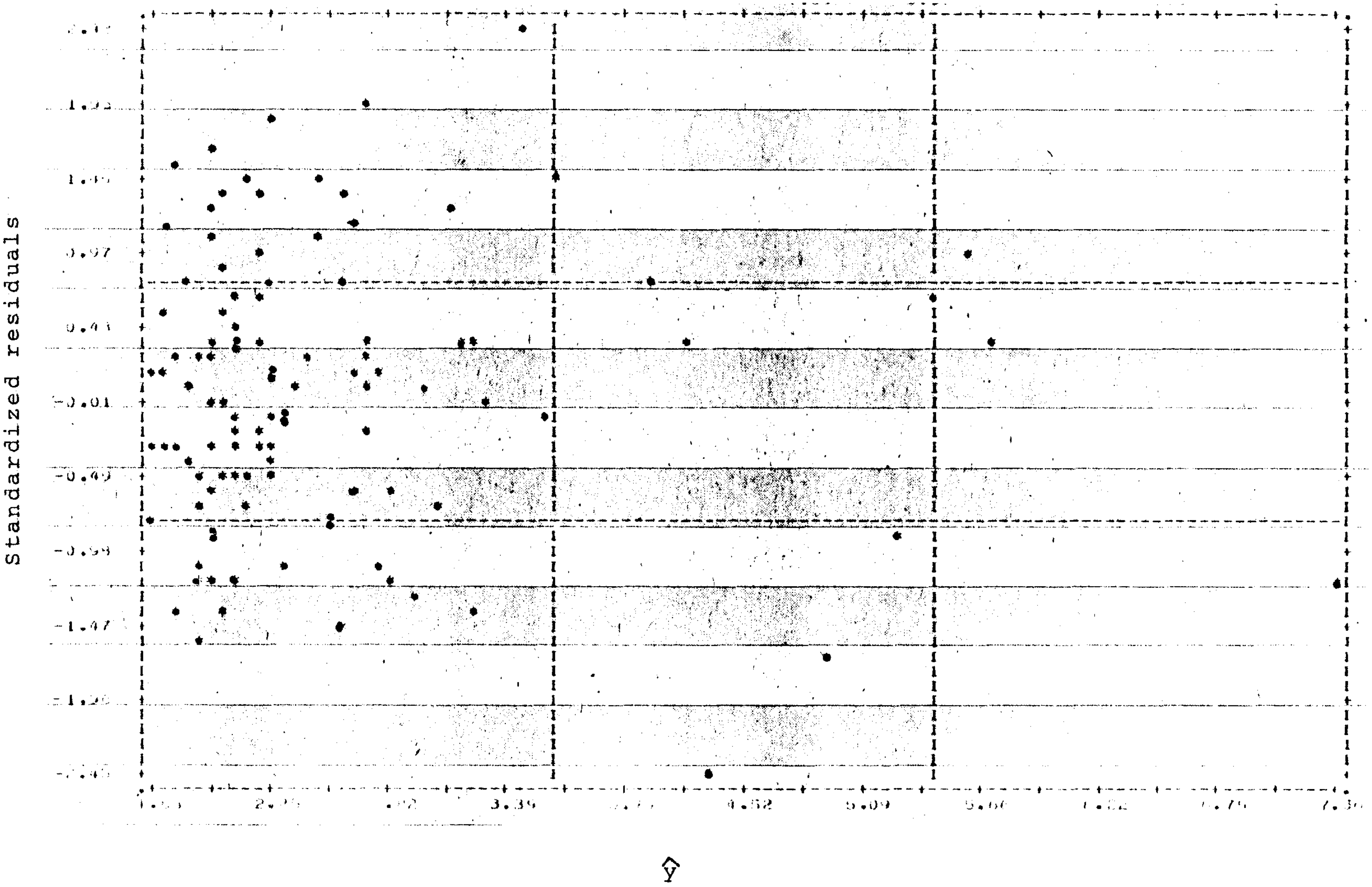


Figure 45. Graph of standardized residuals versus 10SCAV/(36-AGE), transformed regression, female sample.

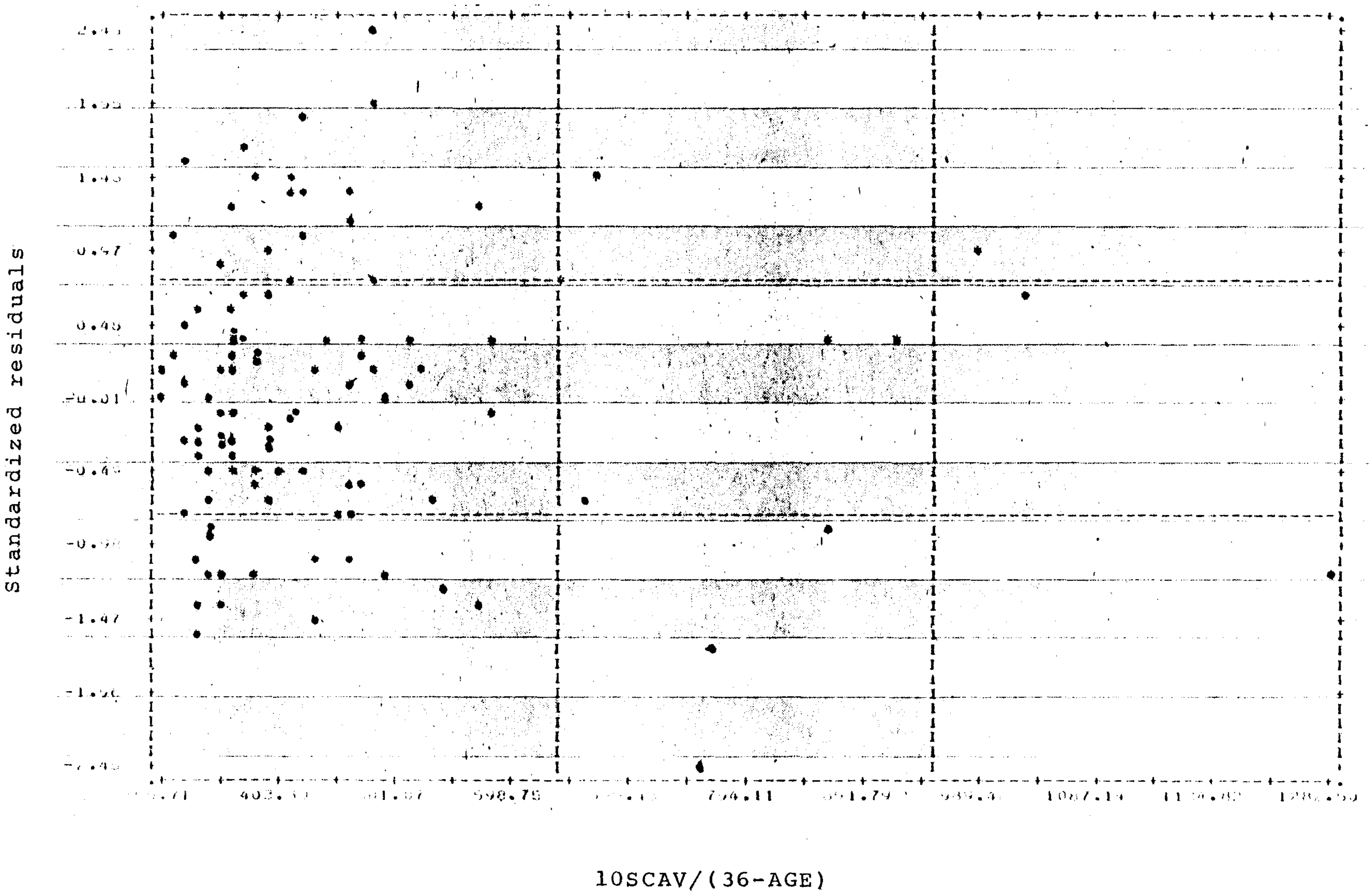


Figure 46. Graph of standardized residuals versus 10YRSB/(36-AGE), transformed regression, female sample.

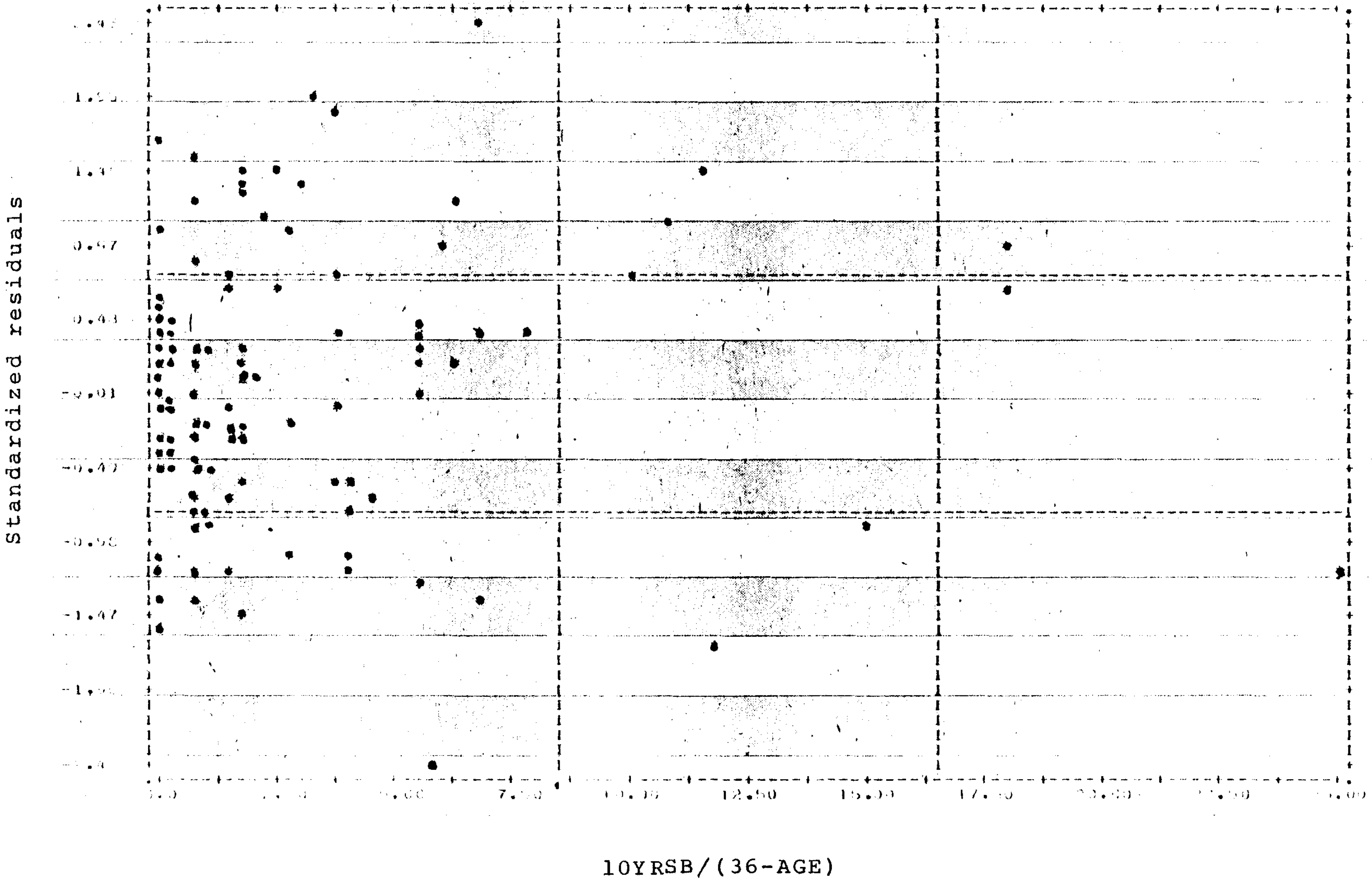


Figure 47. Graph of standardized residuals versus 10QUAN/(36-AGE), transformed regression, female sample.

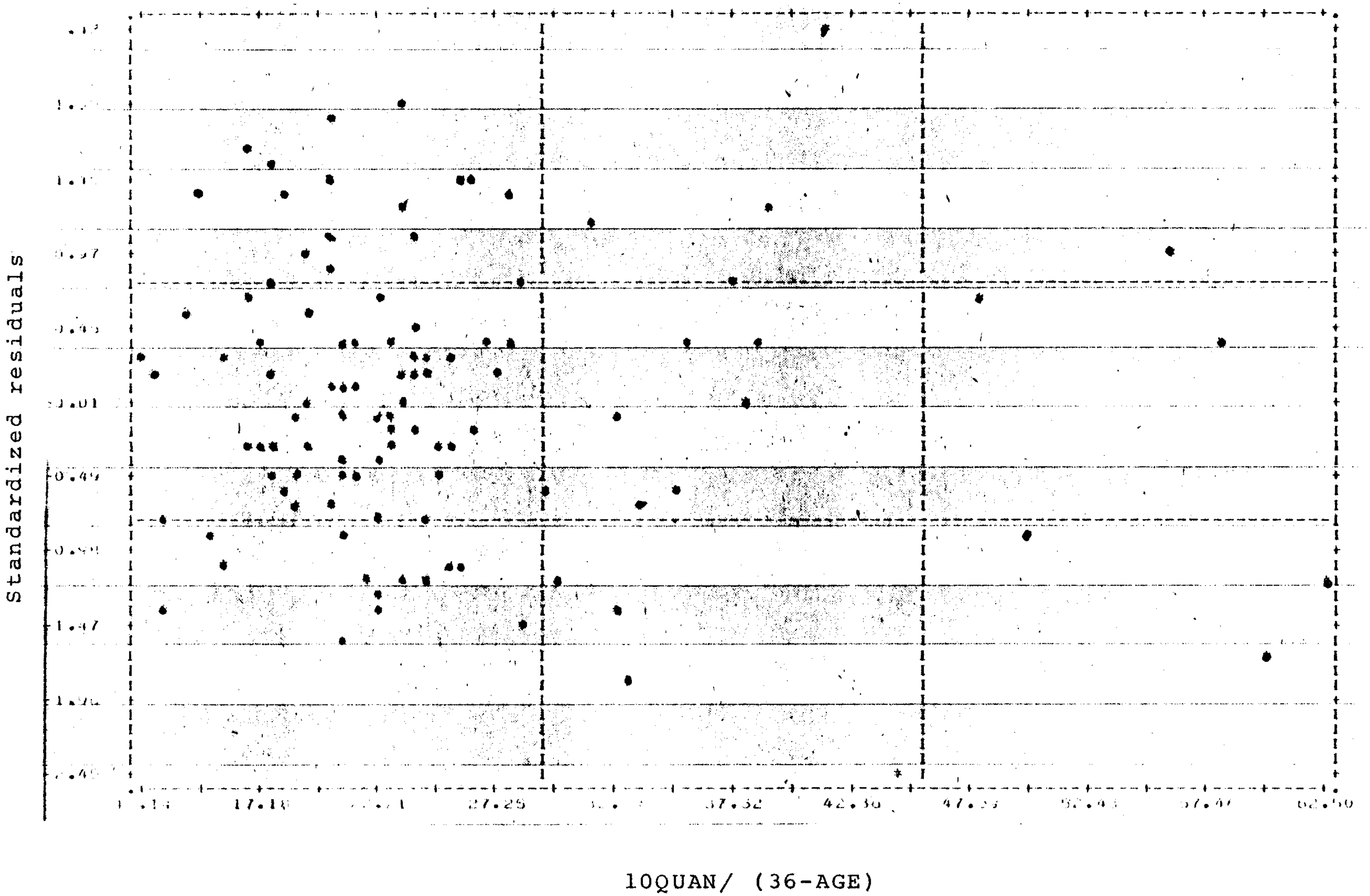


Figure 48. Graph of standardized residuals versus 10NOSC/(36-AGE), transformed regression, female sample.

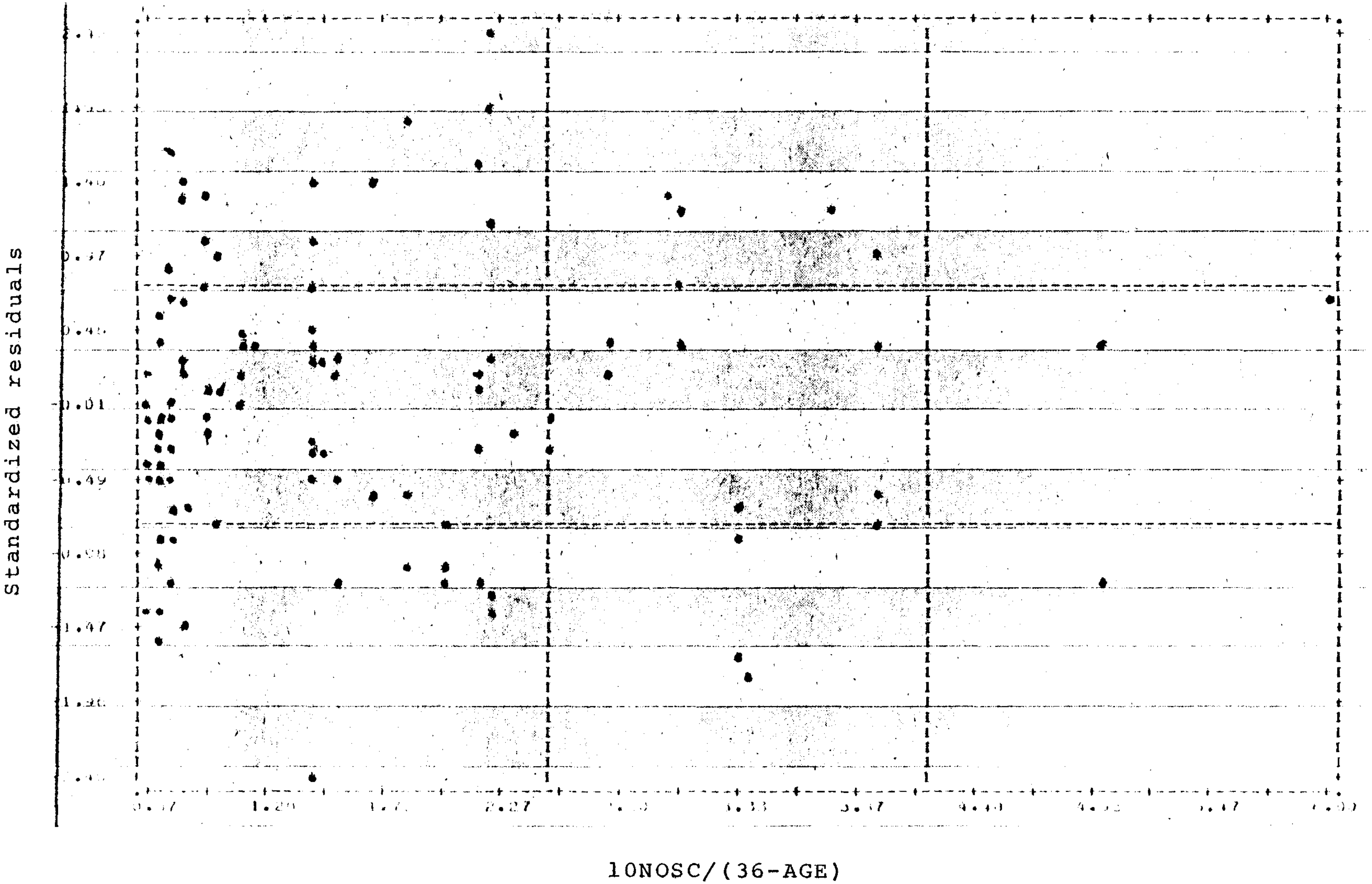


Figure 49. Graph of standardized residuals versus 10UGPA/(36-AGE), transformed regression, female sample.

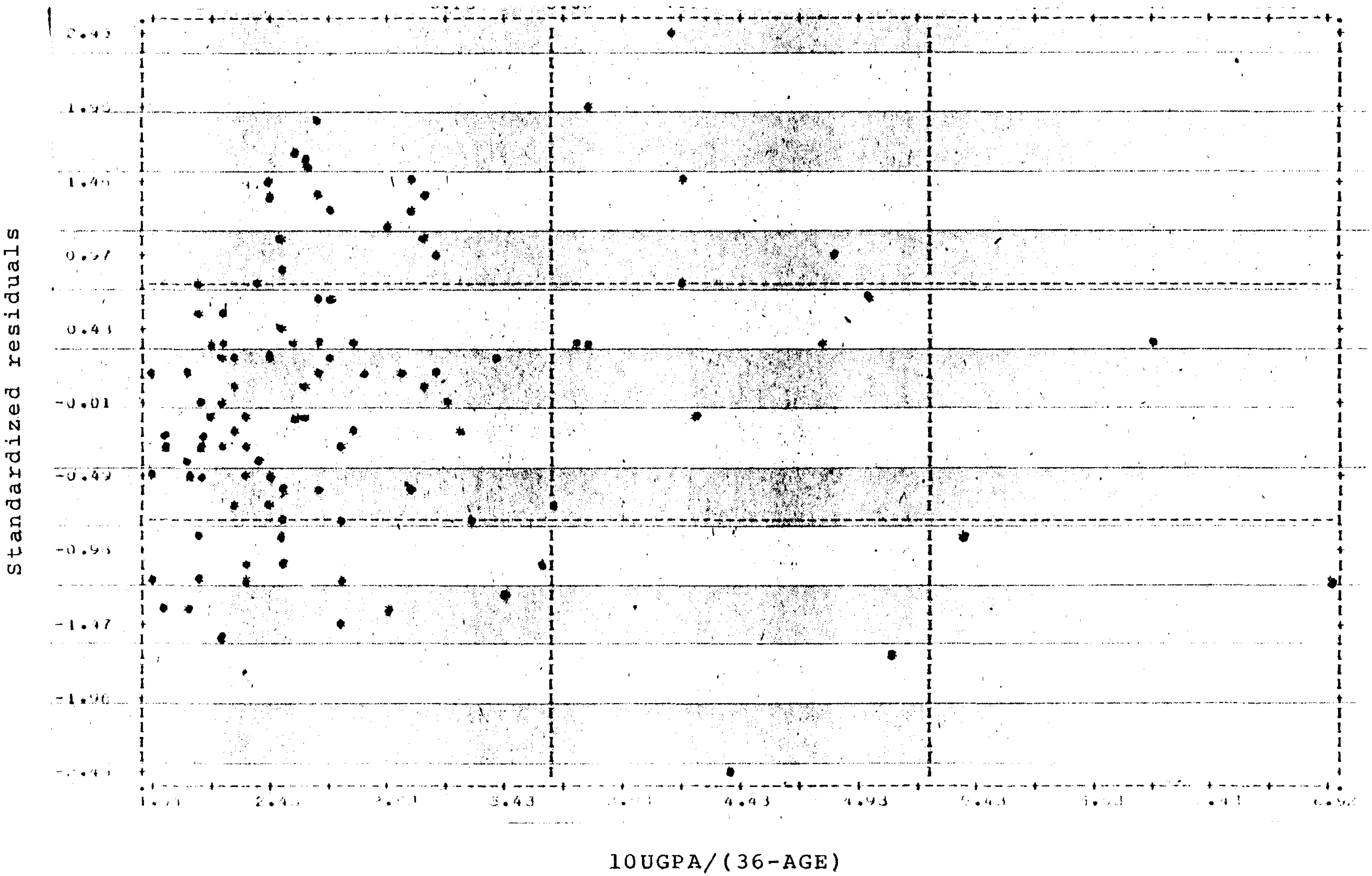


Figure 50. Graph of standardized residuals versus 10WORK/(36-AGE), transformed regression, female sample.

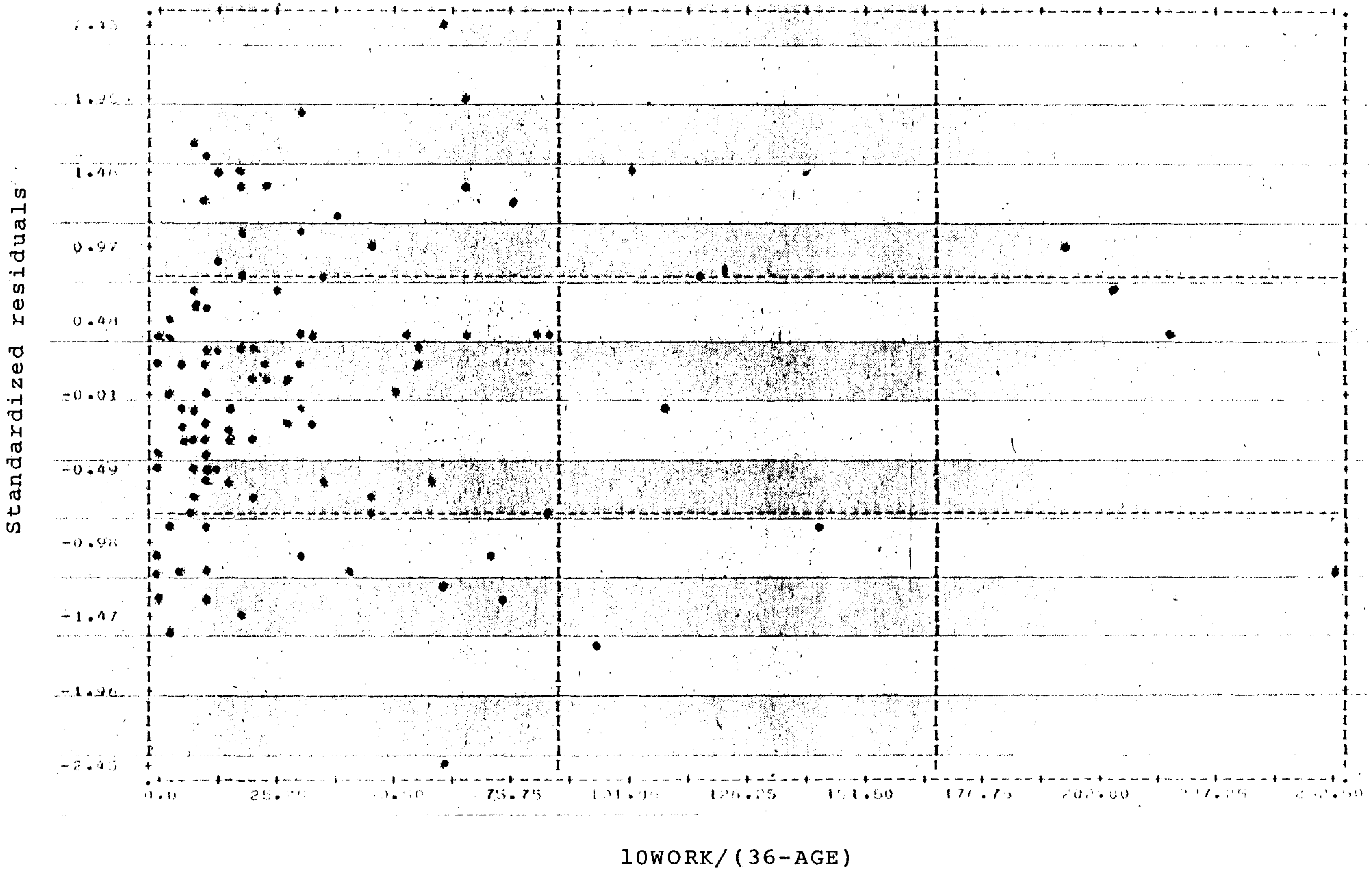


Figure 51. Graph of standardized residuals versus 10VERB/(36-AGE), transformed regression, female sample.

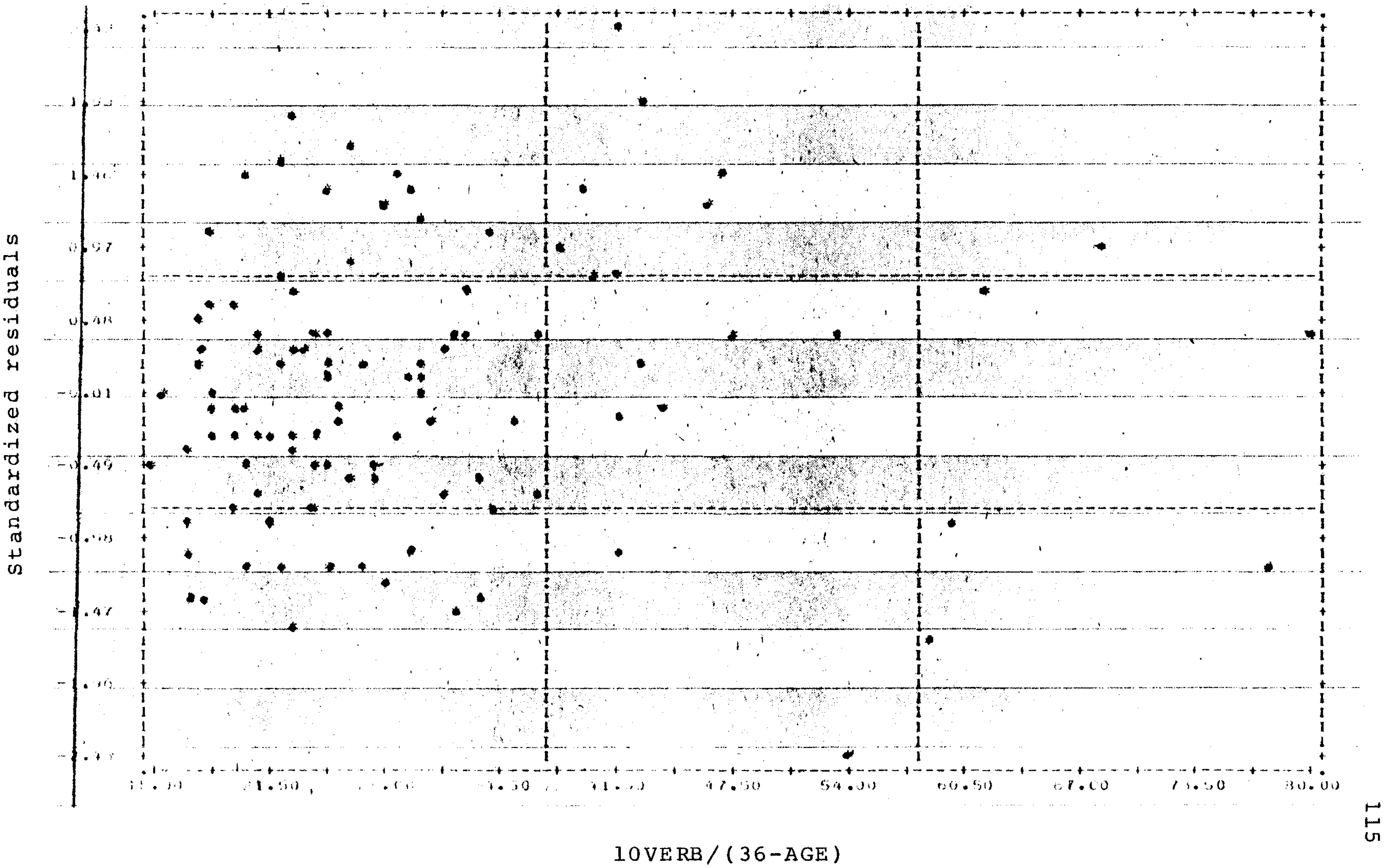
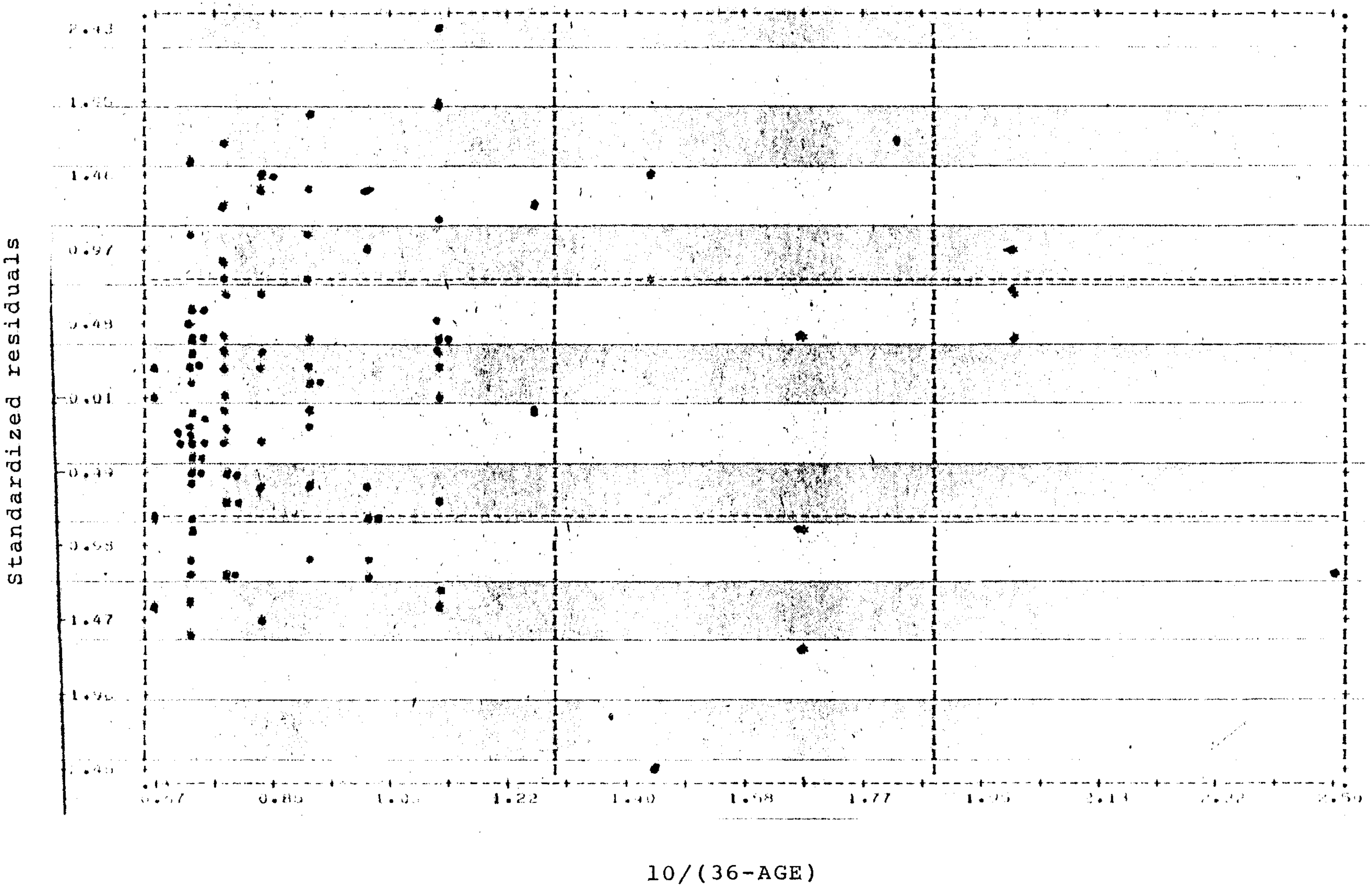


Figure 52. Graph of standardized residuals versus 10/(36-AGE), transformed regression, female sample.



YRSB,WORK and 10/(36-AGE), these points, though few, are well scattered and not indicative of a decreasing funnel shape which showed model violations in earlier plots. The graphs of VERB, UGPA, NOSC and QUAN have only a few points in the last third of the graphs, but again, these show no indication of heteroscedasticity.

Regression and Residual Plots for the Male Sample

The first multiple regression for males used GGPA as the dependent variable and AGE, NOSC, YRSB, WORK, VERB, QUAN, SCAV and UGPA as independent variables. The regression was run in a stepwise fashion using the Statistical Package for the Social Sciences (SPSS) subprogram Regression (Kim and Kohout, 1975). The results from the run are shown in Table 16. The Durbin-Watson statistic was calculated to be 1.838, indicating that autocorrelation was not present.

The final $R^2$ for this run was .2388, indicating that almost 24% of the variance in GGPA was accounted for by variance in the independent variables. F = 12.1915 with (7,272) degrees of freedom implies that $R^2$ is significantly different from zero at both the .05 and .01 levels (Hays, 1973, p. 888). QUAN, UGPA and YRSB contributed 8%, 5% and 6% to $R^2$, respectively. VERB contributed almost 2%. The other variables, SCAV, AGE and WORK each contributed less than 1% after the other variables were entered. NOSC

Table 16. Initial Regression, Male Sample.

| VARIABLE | MULTIPLE R | R SQUARE | B | BETA |
|---|---|---|---|---|
| QUAN | 0.29283 | 0.08575 | 0.016699 | 0.25463 |
| UGPA | 0.37282 | 0.13900 | 0.344942 | 0.31373 |
| YRSB | 0.45062 | 0.20306 | 0.034895 | 0.30181 |
| VERB | 0.47179 | 0.22259 | 0.008667 | 0.13072 |
| SCAV | 0.48112 | 0.23147 | 0.001218 | 0.10742 |
| AGE | 0.48397 | 0.23423 | -0.017677 | -0.18808 |
| WORK | 0.48869 | 0.23882 | 0.001508 | 0.14080 |
| (CONSTANT) | | | 0.724477 | |

| | |
|---|---|
| MULTIPLE R | 0.48869 |
| R SQUARE | 0.23882 |
| STANDARD ERROR | 0.34951 |

| ANALYSIS OF VARIANCE | DF | SUM OF SQUARES | MEAN SQUARE | F |
|---|---|---|---|---|
| REGRESSION | 7 | 10.42506 | 1.48929 | 12.19150 |
| RESIDUAL | 272 | 33.22708 | 0.12216 | |

was not entered into the equation, indicating that it would have added less than .001 to the value of $R^2$. The prediction equation generated by this program was

$$(4) \quad \widehat{GGPA} = .0167QUAN + .3449UGPA + .0349YRSB + .0087VERB + .0012SCAV - .0177AGE + .0015WORK + .7245$$

with a standard error of .3495. The only variable to receive a negative weight was AGE, implying that AGE acted as a suppressor variable in this equation. The largest beta weights were given to QUAN, UGPA and YRSB.

Residuals were calculated and standardized. Plots of the standardized residuals versus each variable are shown in Figures 53 through 60. Inspection of the plots reveals generally a nicely random pattern of points with a few points which can be considered outliers. These points are circled in the residual plots.

As a first step in attempting to improve the prediction for males, these points were removed and the regression was rerun. Three data points were eliminated, two hundred seventy-seven remained. The results from the second regression are shown in Table 17.

The variables entered the equation in the same order as before. However, NOSC entered the second equation though it had not entered the first. $R^2$ was .23690. The prediction equation generated by this program was

$$(5) \quad \widehat{GGPA} = .0175QUAN + .3616UGPA + .0355YRSB +$$

Figure 53. Graph of standardized residuals versus $\hat{y}$, initial regression, male sample.

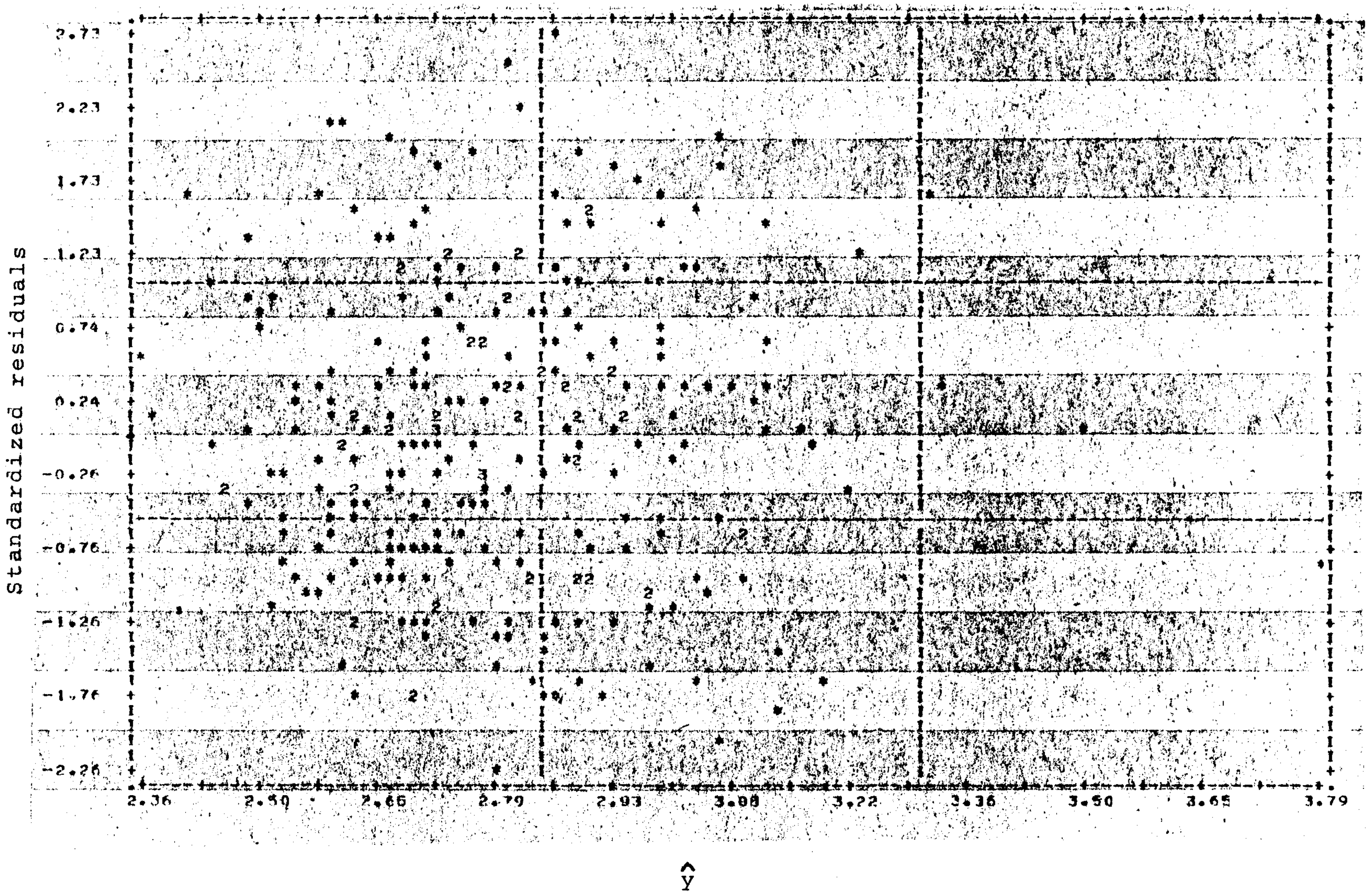


Figure 54. Graph of standardized residuals versus QUAN, initial regression, male sample.

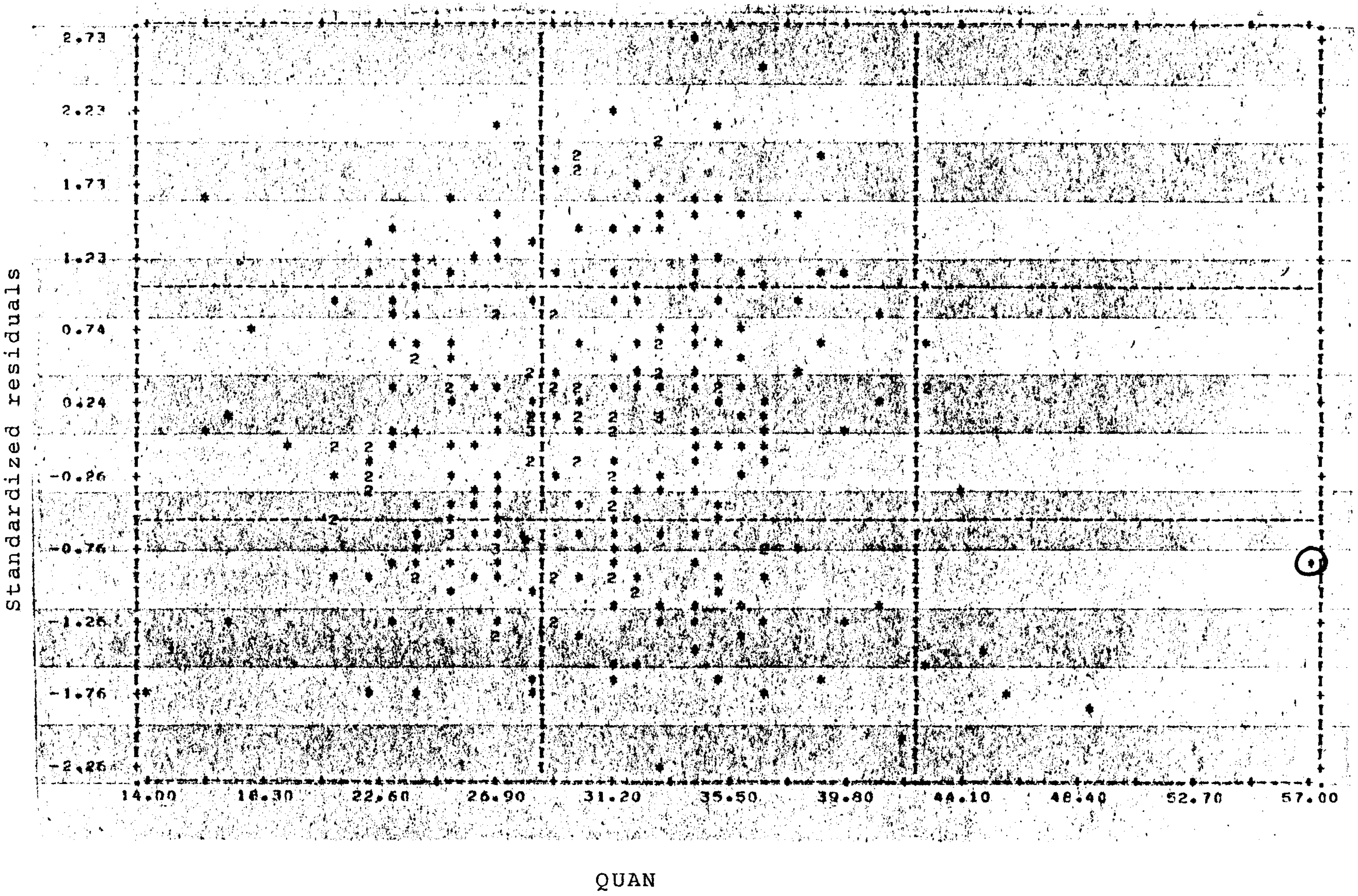


Figure 55. Graph of standardized residuals versus UGPA, initial regression, male sample.

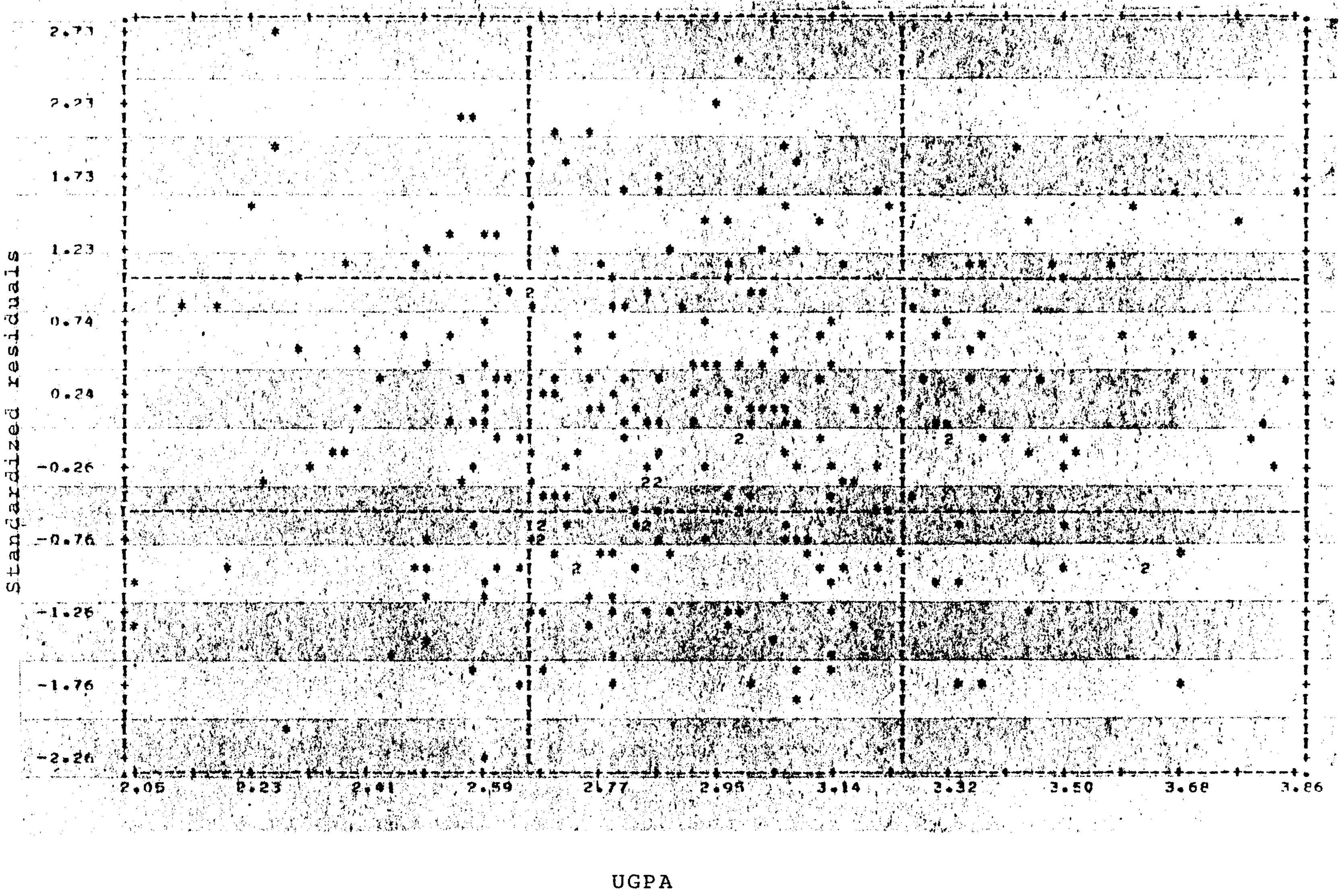


Figure 56. Graph of standardized residuals versus YRSB, initial regression, male sample.

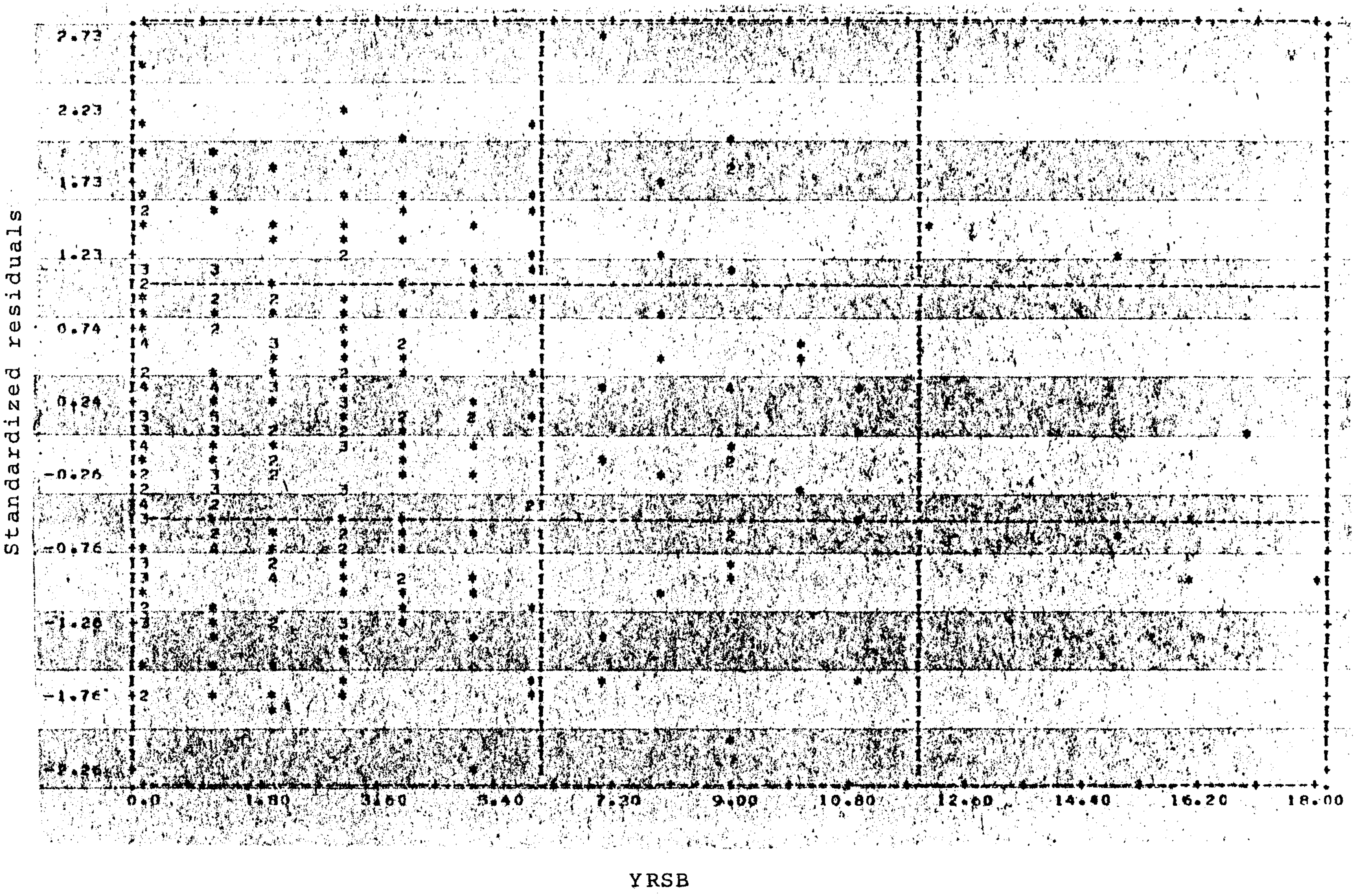


Figure 57. Graph of standardized residuals versus VERB, initial regression, male sample.

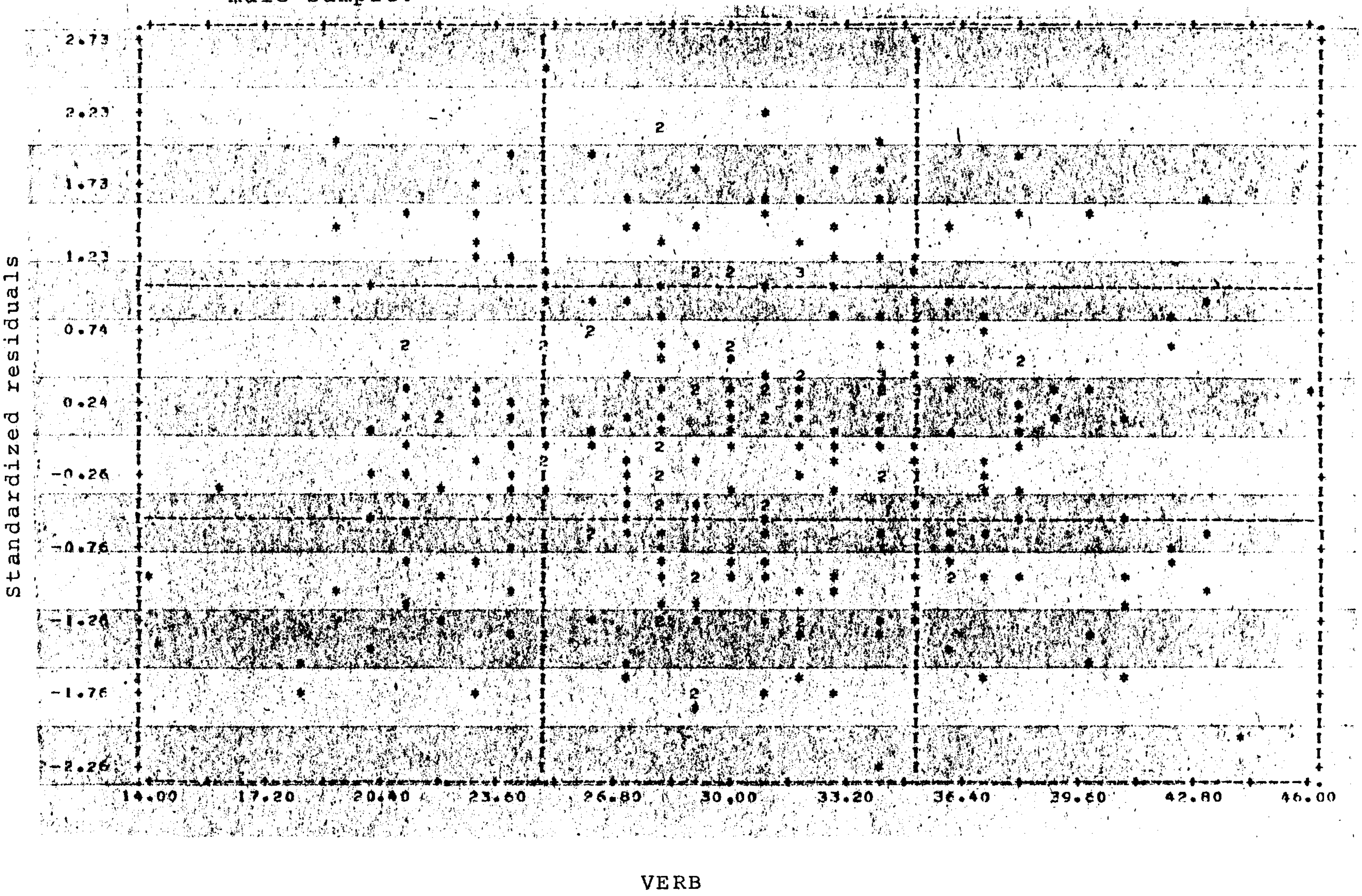


Figure 58. Graph of standardized residuals versus SCAV, initial regression, male sample.

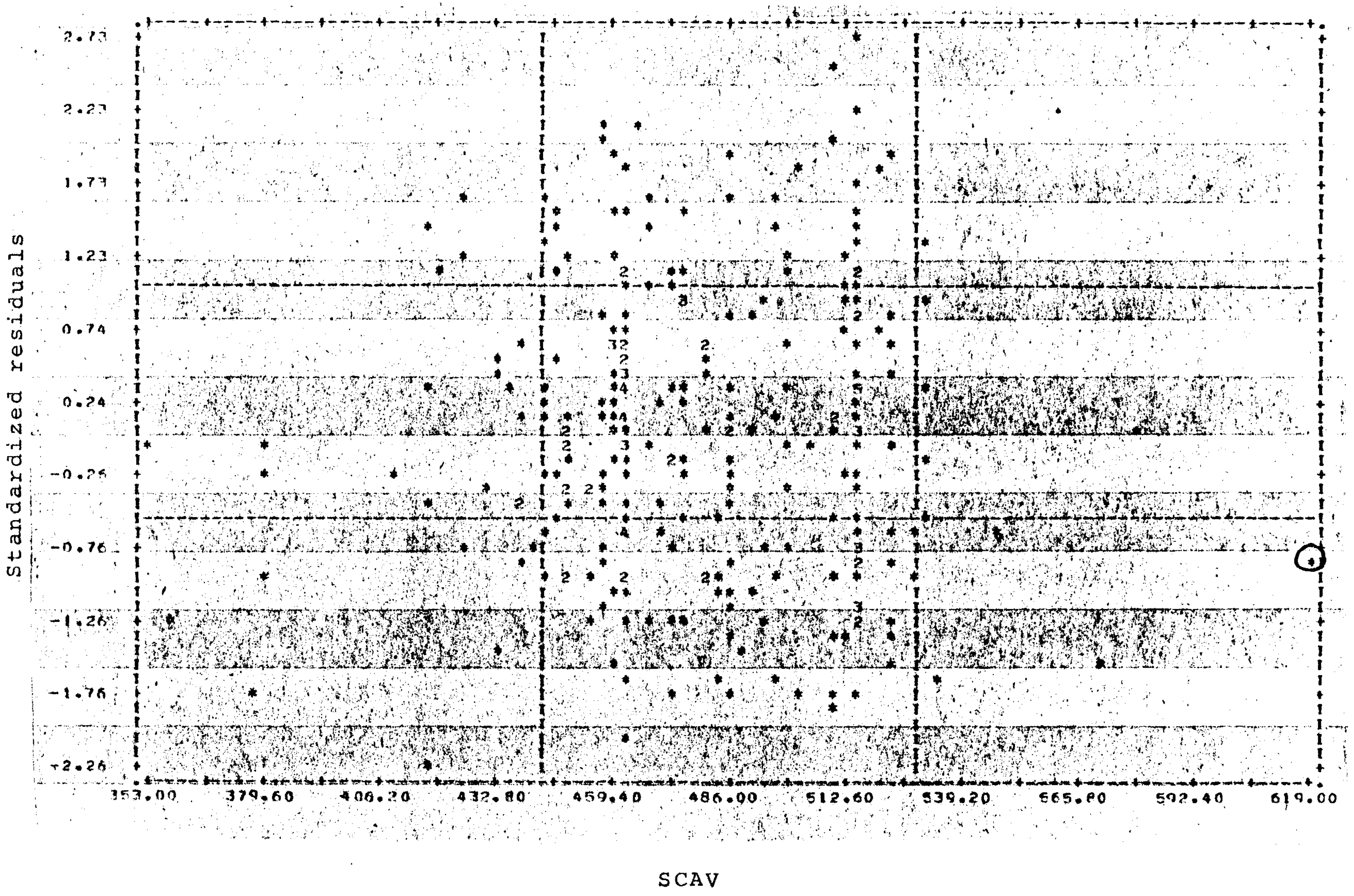


Figure 59. Graph of standardized residuals versus AGE, initial regression, male sample.

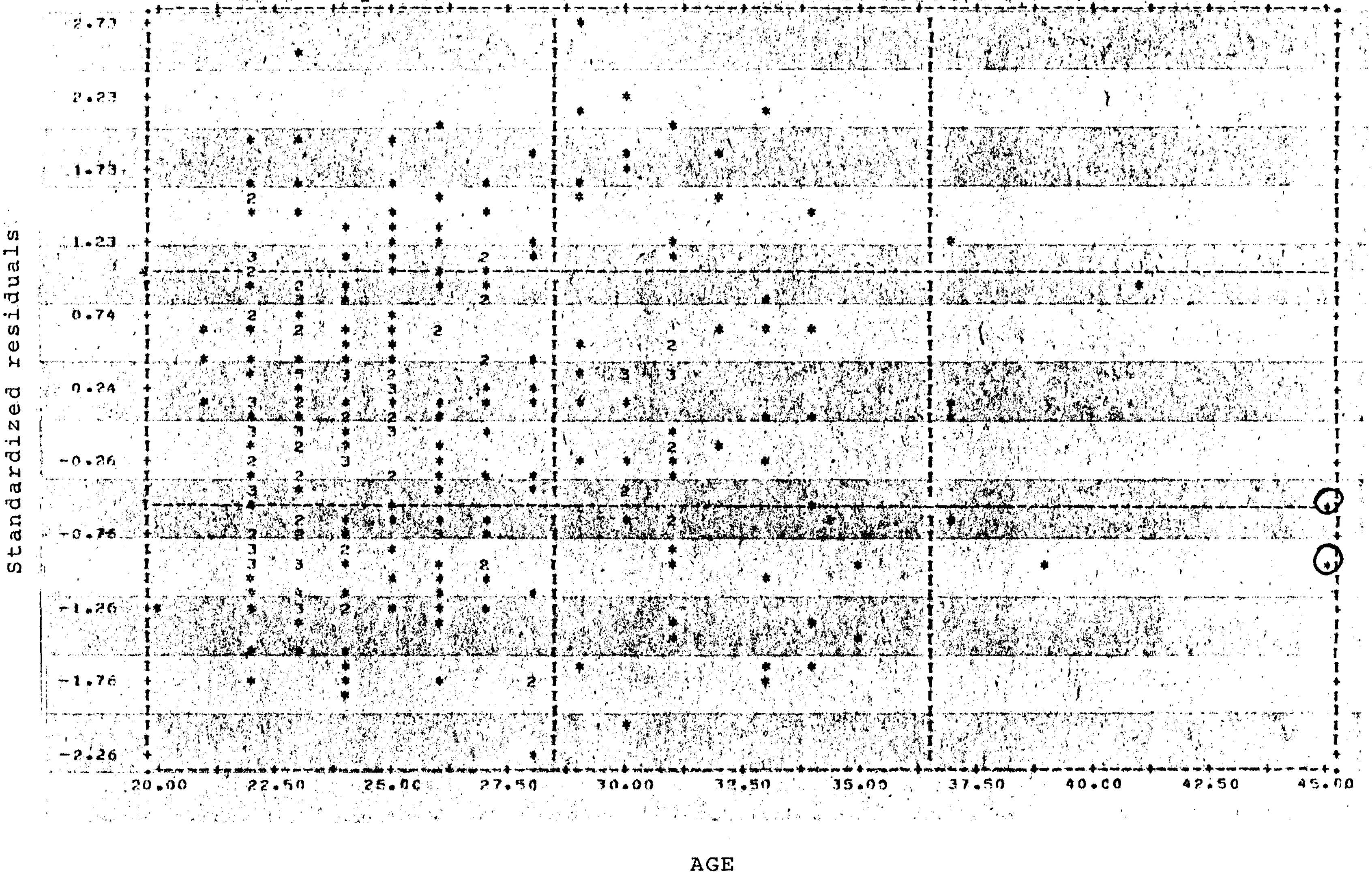


Figure 60. Graph of standardized residuals versus WORK, initial regression, male sample.

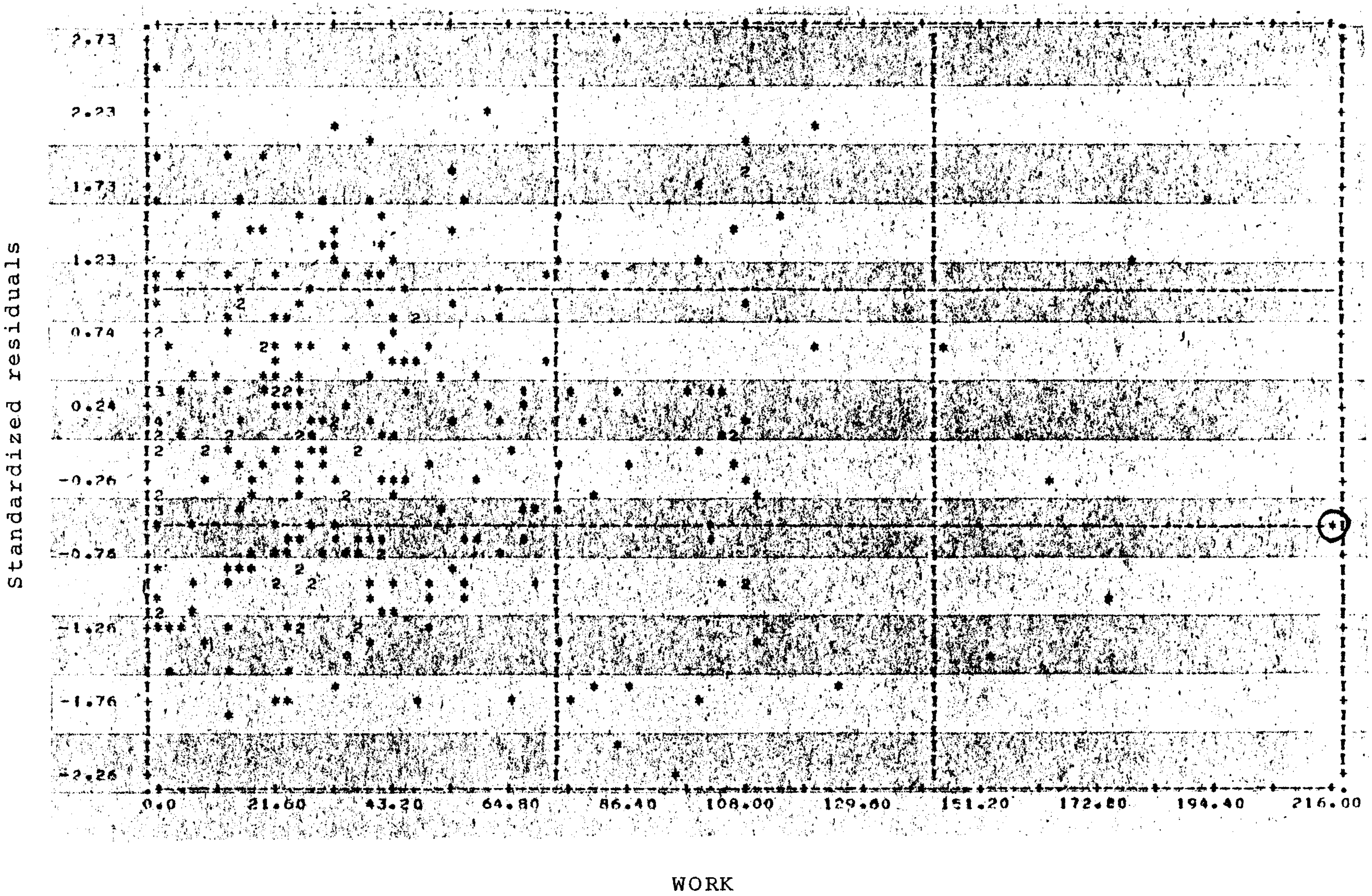


Table 17. Regression with Outliers Removed, Male Sample.

| VARIABLE | MULTIPLE R | R SQUARE | B | BETA |
|---|---|---|---|---|
| QUAN | 0.27474 | 0.07548 | 0.017514 | 0.25851 |
| UGPA | 0.36092 | 0.13026 | 0.361636 | 0.32651 |
| YRSB | 0.45718 | 0.20901 | 0.035456 | 0.29632 |
| VERB | 0.47390 | 0.22458 | 0.008241 | 0.12280 |
| SCAV | 0.48368 | 0.23395 | 0.001230 | 0.10431 |
| AGE | 0.48463 | 0.23487 | -0.112135 | -0.11066 |
| WORK | 0.48666 | 0.23683 | 0.001057 | 0.09460 |
| NOSC | 0.48673 | 0.23690 | -0.004623 | -0.00898 |
| (CONSTANT) | | | 0.517856 | |

| | |
|---|---|
| MULTIPLE R | 0.48673 |
| R SQUARE | 0.23690 |
| STANDARD ERROR | 0.35033 |

| ANALYSIS OF VARIANCE | DF | SUM OF SQUARES | MEAN SQUARE | F |
|---|---|---|---|---|
| REGRESSION | 8 | 10.21125 | 1.27641 | 10.40005 |
| RESIDUAL | 268 | 32.89185 | 0.12273 | |

$$.0082VERB + .0012\ SCAV - .1121AGE +$$
$$.0011WORK - .0046NOSC + .5179$$

with a standard error of .35033. A comparison of the parameter values of the first and second regressions shows that these values are very similar. This is not surprising since only three of two hundred eighty cards were removed.

The scattergrams of the residuals with outliers eliminated are shown in Figures 61 through 69. These scattergrams are well distributed in a random scatter with most of the points falling between plus and minus two. There is no indication of model misspecification or the violation of assumptions. Hence, no transformations would be proper. In several of the plots, specifically, YRSB, SCAV and QUAN, the first or last section of points seems to be slightly underpredicted. In the plots of YRSB and QUAN, more of the points in the last third of the graph are below zero than above. In the graph of SCAV, this pattern of slight underprediction occurs in both the first and last thirds. Hence, we may try adding some additional terms to see if the $R^2$ value can be improved.

## Regressions with Added Terms for the Male Sample

Several other regressions were run with different terms added. In the first of these, the dummy variables of other degree (OTDG) and undergraduate major (DUMM) were included. The results are shown in Table 18. The

Figure 61. Graph of standardized residuals versus $\hat{y}$, regression with outliers removed, male sample.

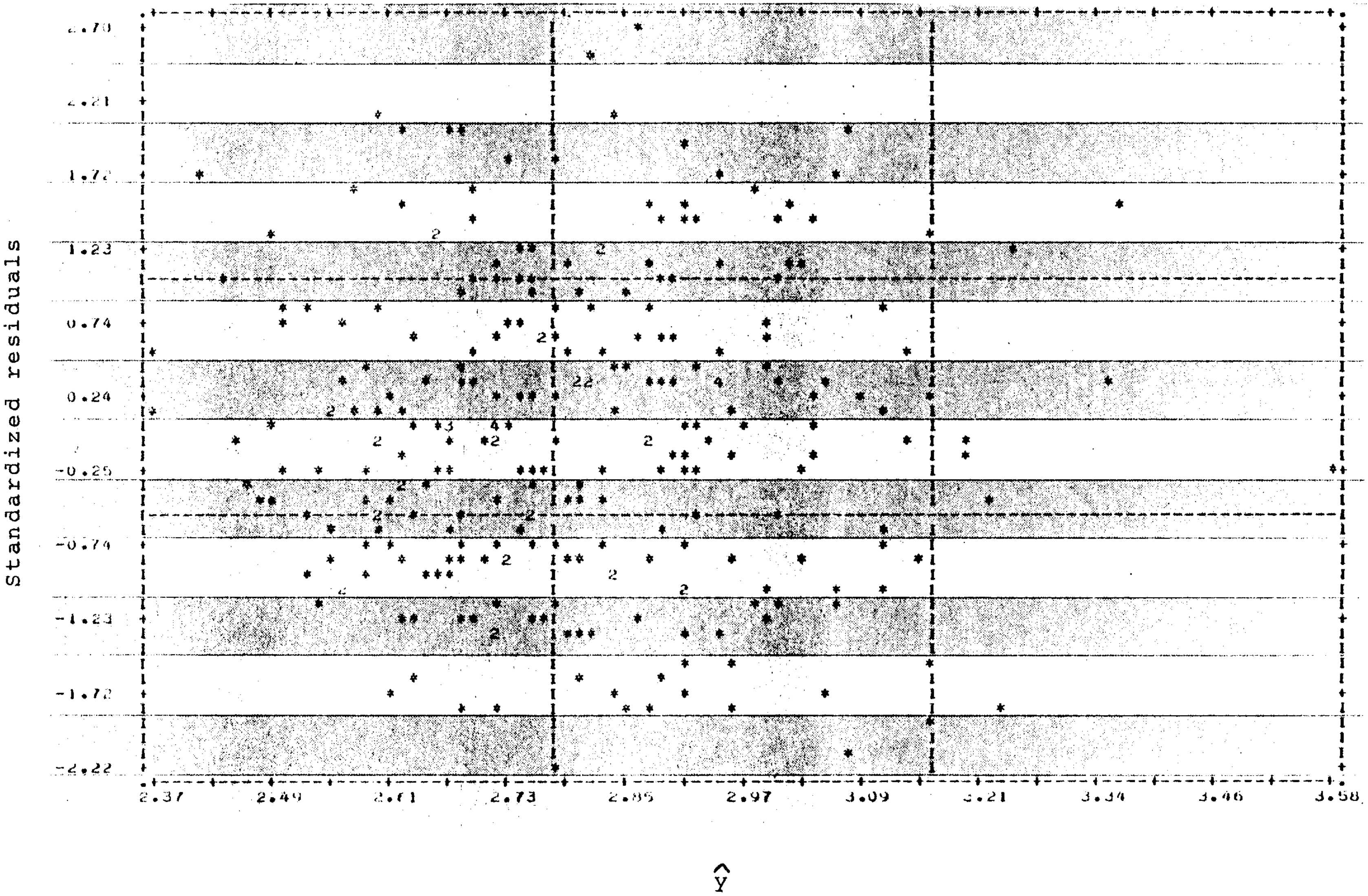


Figure 62. Graph of standardized residuals versus QUAN, regression with outliers removed, male sample.

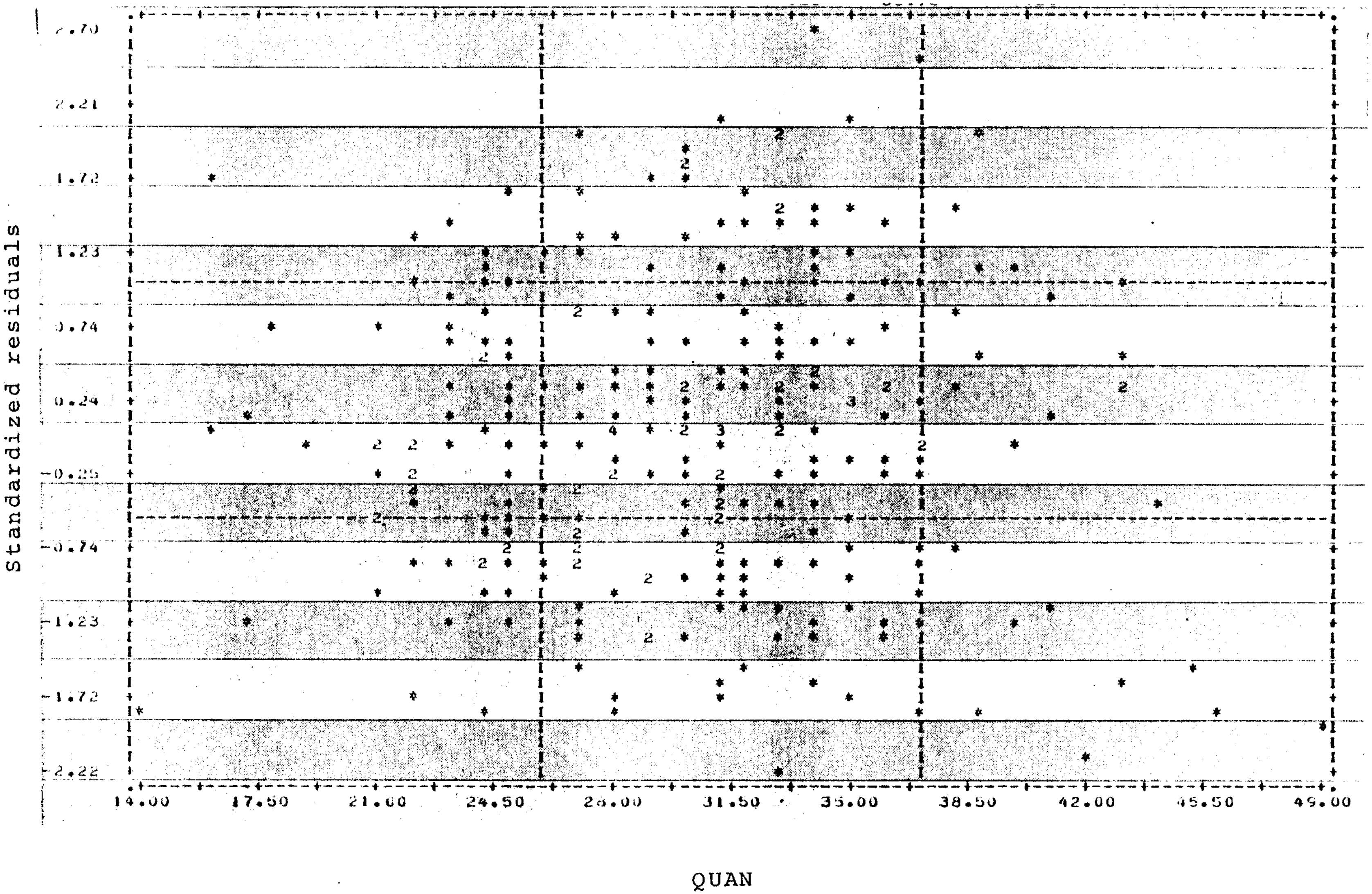


Figure 63. Graph of standardized residuals versus UGPA, regression with outliers removed, male sample.

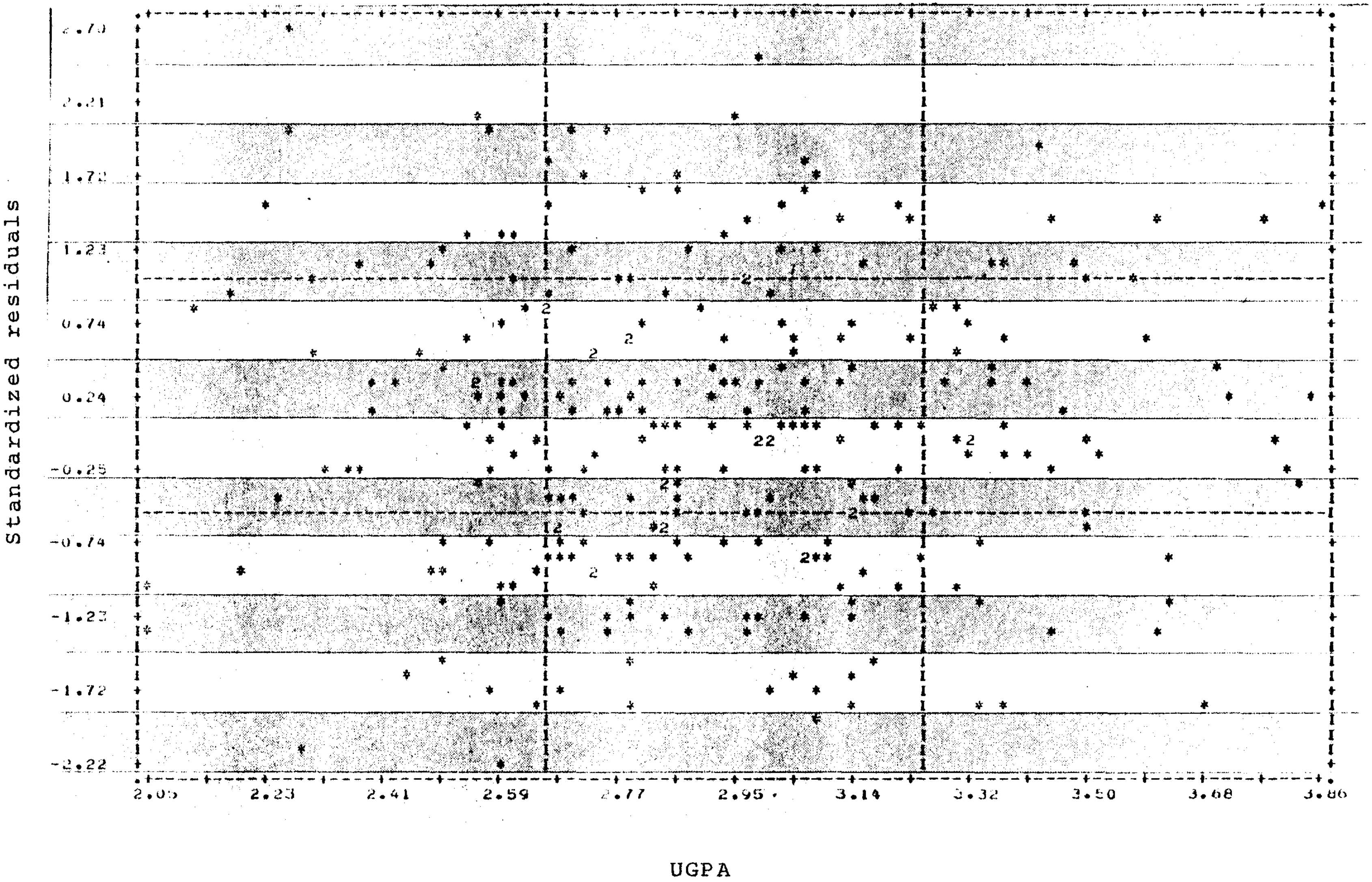


Figure 64. Graph of standardized residuals versus YRSB, regression with outliers removed, male sample.

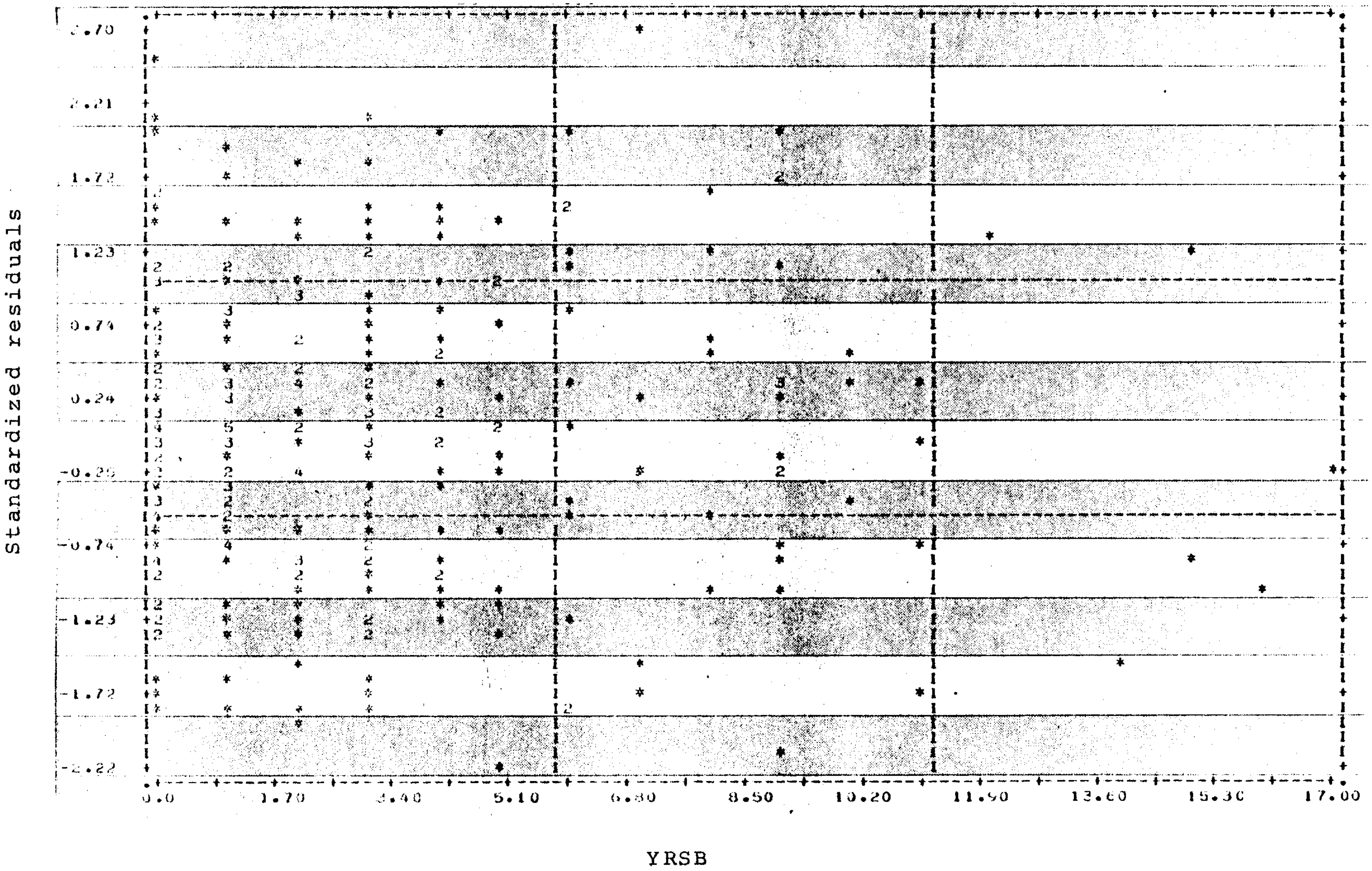


Figure 65. Graph of standardized residuals versus VERB, regression with outliers removed, male sample.

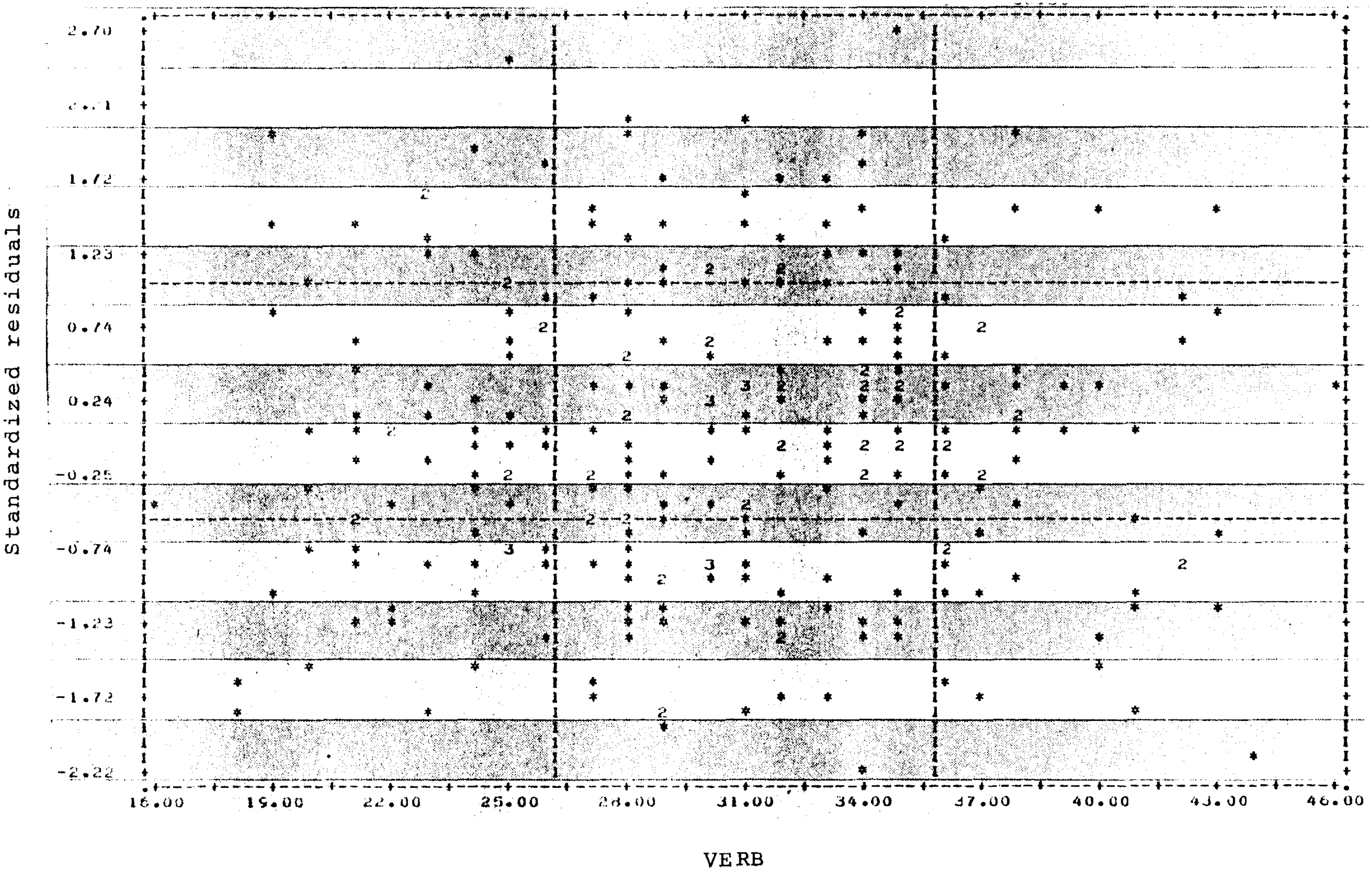


Figure 66. Graph of standardized residuals versus SCAV, regression with outliers removed, male sample.

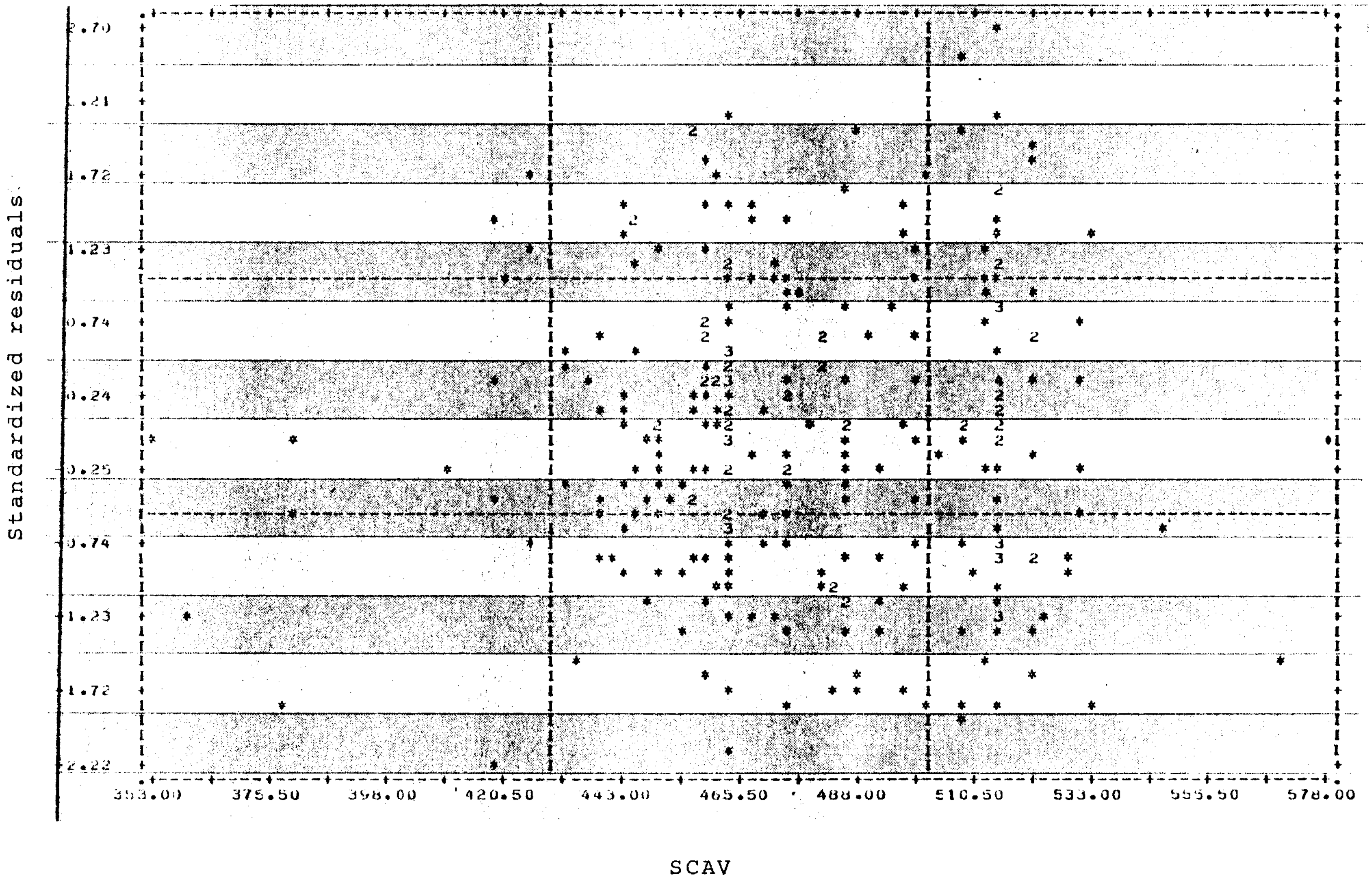


Figure 67. Graph of standardized residuals versus AGE, regression with outliers removed, male sample.

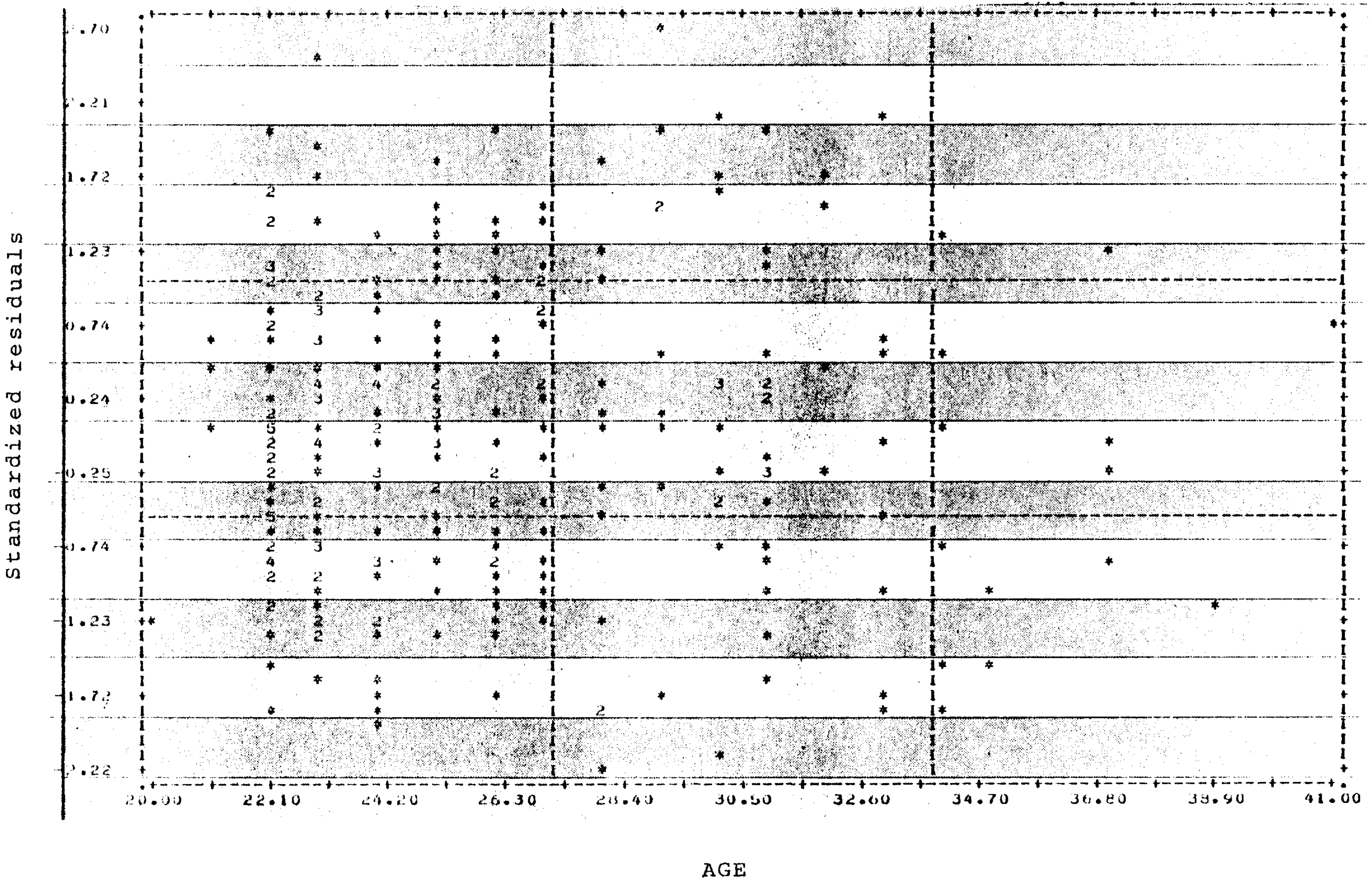


Figure 68. Graph of standardized residuals versus WORK, regression with outliers removed, male sample.

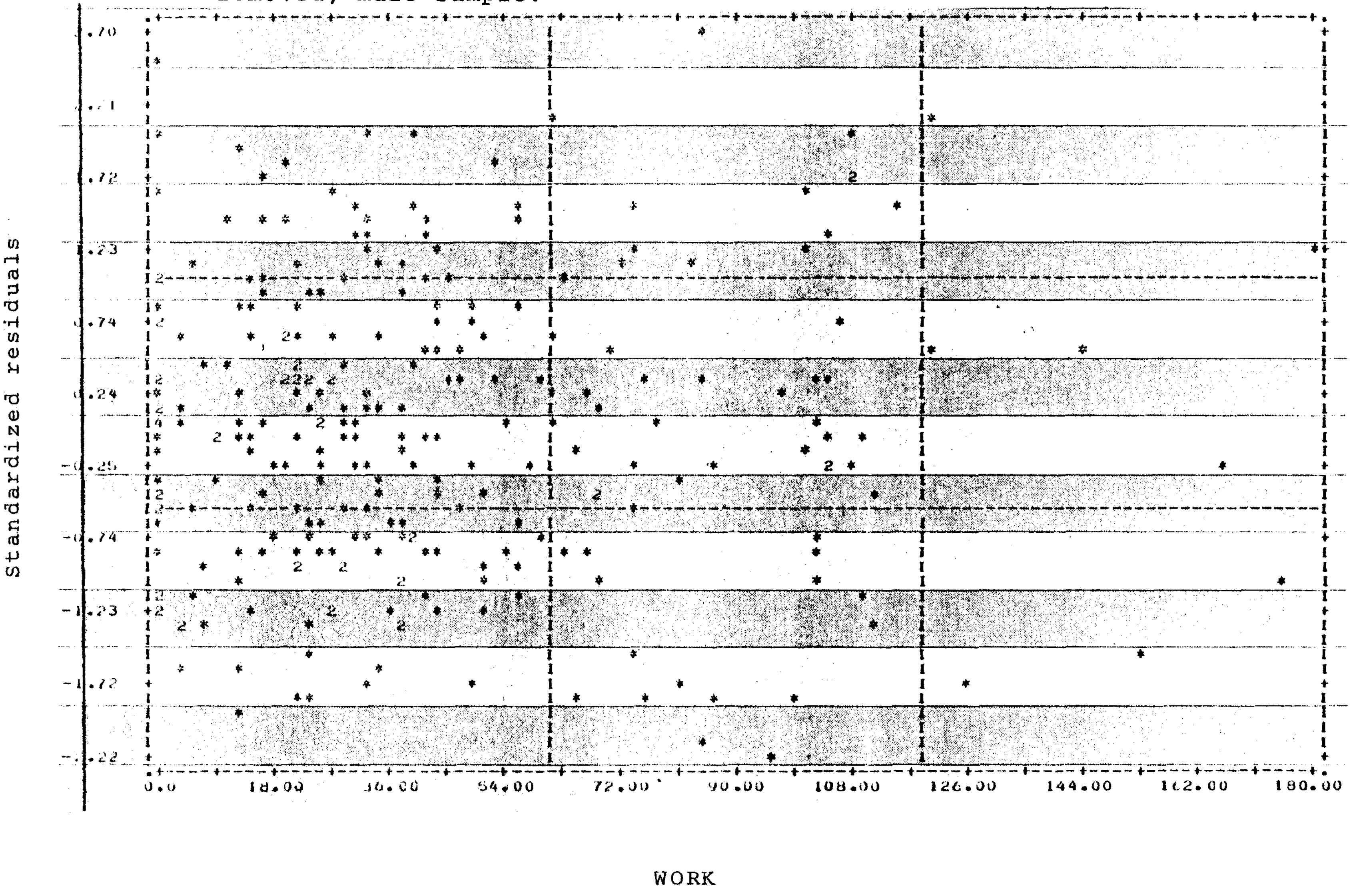


Figure 69. Graph of standardized residuals versus NOSC, regression with outliers removed, male sample.

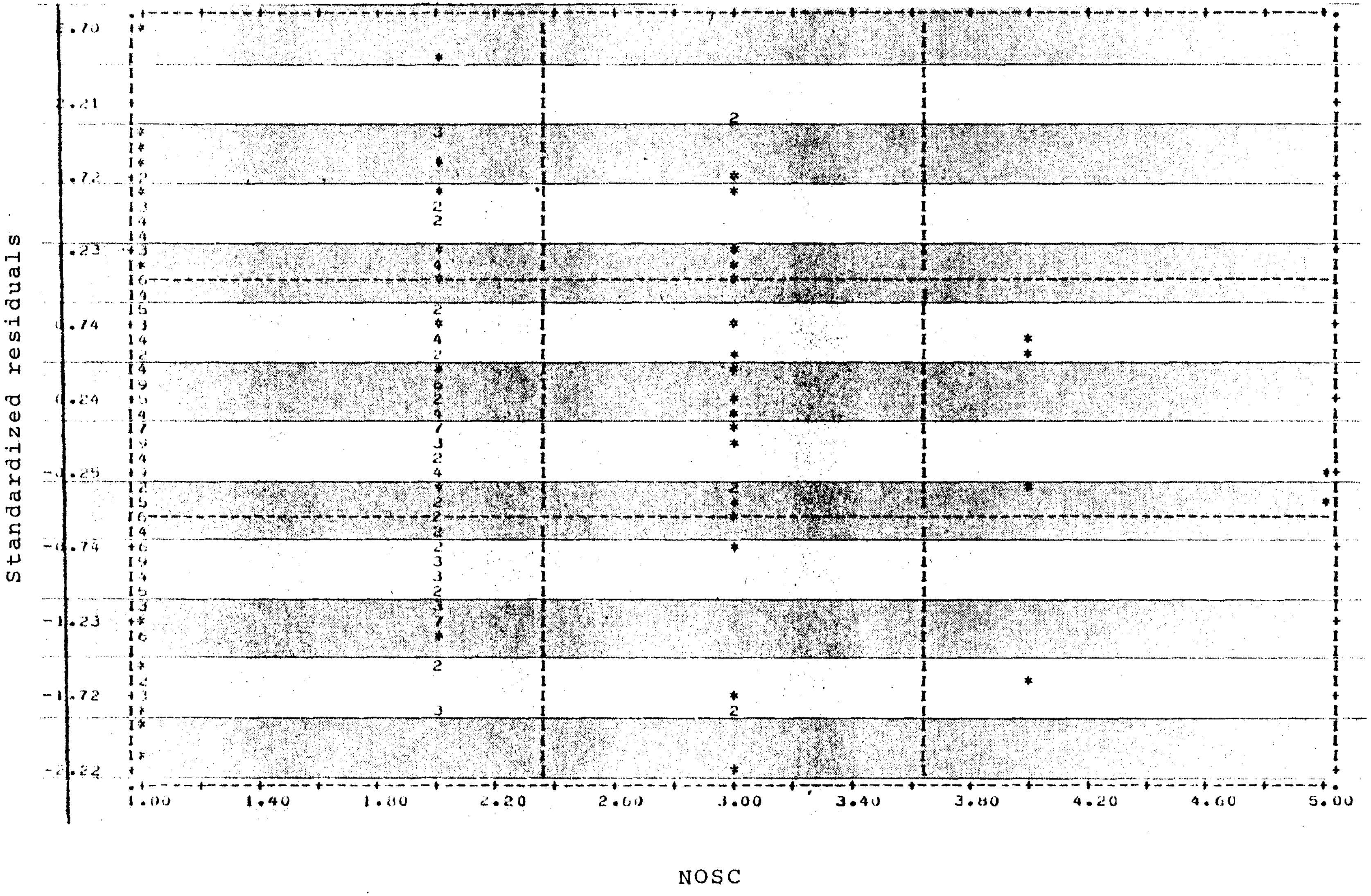


Table 18. Regression with Dummy Variables, Male Sample.

| VARIABLE | MULTIPLE R | R SQUARE | B | BETA |
|---|---|---|---|---|
| QUAN | 0.27474 | 0.07548 | 0.017497 | 0.25826 |
| UGPA | 0.36092 | 0.13026 | 0.360533 | 0.32551 |
| YRSB | 0.45718 | 0.20901 | 0.035744 | 0.29873 |
| VERB | 0.47390 | 0.22458 | 0.008121 | 0.12101 |
| SCAV | 0.48368 | 0.23395 | 0.001267 | 0.10740 |
| SUMM | 0.48563 | 0.23584 | 0.037514 | 0.04755 |
| AGE | 0.48656 | 0.23674 | -0.011561 | -0.11409 |
| WORK | 0.48842 | 0.23855 | 0.001112 | 0.09953 |
| OTDG | 0.48885 | 0.23897 | 0.033922 | 0.02226 |
| NOSC | 0.48906 | 0.23918 | -0.007993 | -0.01552 |
| (CONSTANT) | | | 0.497905 | |

| | |
|---|---|
| MULTIPLE R | 0.48906 |
| R SQUARE | 0.23918 |
| STANDARD ERROR | 0.35112 |

| ANALYSIS OF VARIANCE | DF | SUM OF SQUARES | MEAN SQUARE | F |
|---|---|---|---|---|
| REGRESSION | 10 | 10.30922 | 1.03092 | 8.36209 |
| RESIDUAL | 266 | 32.79388 | 0.12329 | |

$R^2$ value was .23918 and the standard error was .35112. Neither of the added terms contributed much to the equation, either in the $R^2$ change or in the value of the beta weight. DUMM contributed .0019 to the value of $R^2$ and OTDG contributed .0004. Their beta weights were .048 and .022, respectively. The value of $R^2$ increased from .2369 to .2392, or, .0023, as a result of the addition of these two variables.

For the second group of additional terms, five interactions were included. These were UGPA*QUAN, SCAV* QUAN, VERB*QUAN, AGE*WORK and AGE*YRSB. The results of this regression are shown in Table 19. $R^2$ was .25724 and the standard error was .34890. Three of the interactions have beta weights higher in absolute value than .2. These are UGPA*QUAN, AGE*WORK and AGE*YRSB. The amounts contributed to the value of $R^2$ by these variables were .137, .001 and .006, respectively. The other two interactions, QUAN*SCAV and QUAN*VERB, had beta weights of -.12 and -.08 and each contributed less than .0002 to changing the value of $R^2$. AGE*YRSB received a negative weight, indicating that it acted as a suppressor variable. This is due to AGE since AGE has acted as a suppressor variable in all the previous programs. The value of $R^2$ increased from .2390 to .2572, or, .0203, as a result of the addition of these five interaction terms.

The next two regressions included terms which were

Table 19. Regression with Interaction Terms, Male Sample.

| VARIABLE | MULTIPLE R | R SQUARE | B | BETA |
|---|---|---|---|---|
| QUAN*UGPA | 0.36974 | 0.13670 | 0.018956 | 0.90280 |
| YRSB | 0.45681 | 0.20867 | 0.169410 | 1.41583 |
| VERB | 0.46926 | 0.22020 | 0.011499 | 0.17137 |
| QUAN | 0.48386 | 0.23412 | -0.026898 | -0.39703 |
| SCAV | 0.49274 | 0.24280 | 0.001773 | 0.15031 |
| AGE*YRSB | 0.49925 | 0.24925 | -0.004101 | -1.13926 |
| AGE*WORK | 0.50036 | 0.25036 | 0.000254 | 0.76985 |
| AGE | 0.50315 | 0.25316 | -0.020407 | -0.20139 |
| WORK | 0.50564 | 0.25568 | -0.006247 | -0.55896 |
| UGPA | 0.50690 | 0.25695 | -0.193017 | -0.17427 |
| QUAN*VERB | 0.50703 | 0.25708 | -0.000122 | -0.08745 |
| NOSC | 0.50713 | 0.25718 | 0.005225 | 0.01014 |
| QUAN*SCAV | 0.50719 | 0.25724 | -0.000016 | -0.12848 |
| (CONSTANT) | | | 1.975765 | |

| | |
|---|---|
| MULTIPLE R | 0.50719 |
| R SQUARE | 0.25724 |
| STANDARD ERROR | 0.34890 |

| ANALYSIS OF VARIANCE | DF | SUM OF SQUARES | MEAN SQUARE | F |
|---|---|---|---|---|
| REGRESSION | 13 | 11.08798 | 0.85292 | 7.00664 |
| RESIDUAL | 263 | 32.01512 | 0.12173 | |

squares of variables. In the first case these were AGE*AGE, WORK*WORK and YRSB*YRSB; in the second, they were QUAN*QUAN and SCAV*SCAV. Results from these two regressions are shown in Tables 20 and 21, respectively. The squares of AGE, WORK and YRSB all entered the equation though with minor contributions to the value of $R^2$. AGE*AGE, however, had a large negative beta weight. $R^2$ decreased slightly from the previous program to .24163 and the standard error increased slightly to .35121. With the addition of QUAN*QUAN and SCAV*SCAV, $R^2$ was .25960 and the standard error was .34638. Both squares received negative weights and their beta weights were relatively large in absolute value compared to the other weights. As a result of adding the three squared terms, $R^2$ changed from .2369 in the program with outliers removed to .2416, an increase of .0047. Adding the two squared terms increased $R^2$ by .0227 over the value in Table 17.

## Regression with the Best Variables, Male Sample

Before running a final equation, the F values of all the variables were examined to see which variables might be eliminated. A non-significant F value indicates a variable which can be deleted from the regression equation. The F ratio used by SPSS for each variable, $x_i$, is

$$F = \frac{\text{incremental SS due to } x_i/1}{SS_{res}/(N-k-1)}$$

where F has (1,N-k-1) degrees of freedom (Kim and Kohout,

Table 20. Regression with Three Squared Terms, Male Sample.

| VARIABLE | MULTIPLE R | R SQUARE | B | BETA |
|---|---|---|---|---|
| QUAN | 0.27474 | 0.07548 | 0.017046 | 0.25161 |
| UGPA | 0.36092 | 0.13026 | 0.381653 | 0.34458 |
| YRSB | 0.45718 | 0.20901 | 0.041972 | 0.35078 |
| VERB | 0.47390 | 0.22458 | 0.008387 | 0.12498 |
| SCAV | 0.48368 | 0.23395 | 0.001391 | 0.11795 |
| YRSB*YRSB | 0.48549 | 0.23570 | -0.000699 | -0.07144 |
| AGE*AGE | 0.48663 | 0.23681 | -0.001389 | -0.77250 |
| WORK*WORK | 0.48998 | 0.24008 | 0.000011 | 0.13585 |
| AGE | 0.49128 | 0.24136 | 0.070081 | 0.69158 |
| WORK | 0.49143 | 0.24150 | -0.000560 | -0.05015 |
| NOSC | 0.49156 | 0.24163 | -0.006373 | -0.01237 |
| (CONSTANT) | | | -0.730316 | |

| | |
|---|---|
| MULTIPLE R | 0.49156 |
| R SQUARE | 0.24163 |
| STANDARD ERROR | 0.35121 |

| ANALYSIS OF VARIANCE | DF | SUM OF SQUARES | MEAN SQUARE | F |
|---|---|---|---|---|
| REGRESSION | 11 | 10,41492 | 0.94681 | 7.67570 |
| RESIDUAL | 265 | 32.68818 | 0.12335 | |

Table 21. Regression with Two Squared Terms, Male Sample.

| VARIABLE | MULTIPLE R | R SQUARE | B | BETA |
|---|---|---|---|---|
| QUAN | 0.27474 | 0.07548 | 0.077361 | 1.14190 |
| UGPA | 0.36092 | 0.13026 | 0.374262 | 0.33791 |
| YRSB | 0.45718 | 0.20901 | 0.034790 | 0.29076 |
| VERB | 0.47390 | 0.22458 | 0.008555 | 0.12749 |
| QUAN*QUAN | 0.48979 | 0.23990 | -0.000981 | -0.89700 |
| SCAV | 0.49979 | 0.24979 | 0.017691 | 1.49984 |
| SCAV*SCAV | 0.50590 | 0.25594 | -0.000017 | -1.38789 |
| AGE | 0.50641 | 0.25645 | -0.012614 | -0.12447 |
| WORK | 0.50933 | 0.25942 | 0.001267 | 0.11339 |
| NOSC | 0.50951 | 0.25960 | 0.007511 | 0.01458 |
| (CONSTANT) | | | -4.293488 | |

| | |
|---|---|
| MULTIPLE R | 0.50951 |
| R SQUARE | 0.25960 |
| STANDARD ERROR | 0.34638 |

| ANALYSIS OF VARIANCE | DF | SUM OF SQUARES | MEAN SQUARE | F |
|---|---|---|---|---|
| REGRESSION | 10 | 11.18946 | 1.11895 | 9.32641 |
| RESIDUAL | 266 | 31.91364 | 0.11998 | |

1975, pp. 336-7). At the .05 level, the critical value of F for the male sample is 3.84 (Hays, 1973, p. 888). Table 22 lists the variables in each regression, from Tables 16 through 21, and the value of F as calculated for each variable in each program. Variables with significant F values are designated by an asterisk by their name. Those variables which contributed amounts to $R^2$ significantly different from zero, at the .05 level, were QUAN, UGPA, VERB, YRSB, SCAV and QUAN*QUAN.

These six variables were entered into another regression equation, the remaining variables were eliminated. The results from this regression are shown in Table 23. $R^2$ had a value of .24979. The F value was 14.98 with (6,270) degrees of freedom, hence $R^2$ is significantly different from zero at both the .05 and .01 levels (Hays, 1973, pp. 888-890). All six variables entered the equation and the prediction equation generated by this run was

$$(6)\quad \widehat{\text{GGPA}} = .0799\text{QUAN} + .3535\text{UGPA} + .0323\text{YRSB} + \\ .0086\text{VERB} - .0010\text{QUAN*QUAN} + .0012\text{SCAV} \\ - .6318$$

with a standard error of .34607. QUAN*QUAN received a negative weight, all other weights were positive. The standardized residuals were calculated and plotted versus $\hat{y}$ and each independent variable. These residuals graphs are shown in Figures 70 through 76. The graphs exhibit a nice random scatter. Four of the two hundred seventy-seven

Table 22. Variables and Their F Values, Male Sample.

| TABLE | VARIABLE | F VALUE |
|---|---|---|
| 16 | *QUAN | 20.55 |
| | *UGPA | 30.83 |
| | *YRSB | 14.93 |
| | *VERB | 5.67 |
| | SCAV | 3.68 |
| | AGE | 2.61 |
| | WORK | 1.64 |
| 17 | *QUAN | 21.22 |
| | *UGPA | 32.45 |
| | *YRSB | 12.45 |
| | *VERB | 4.95 |
| | SCAV | 3.39 |
| | AGE | 0.90 |
| | WORK | 0.71 |
| | NOSC | 0.02 |
| 18 | *QUAN | 21.01 |
| | *UGPA | 31.17 |
| | *YRSB | 12.32 |
| | *VERB | 4.77 |
| | SCAV | 3.56 |
| | DUMM | 0.72 |
| | AGE | 0.91 |
| | WORK | 0.75 |
| | OTDG | 0.15 |
| | NOSC | 0.07 |
| 19 | UGPA*QUAN | 3.46 |
| | *YRSB | 5.05 |
| | VERB | 0.32 |
| | QUAN | 0.19 |
| | SCAV | 0.32 |
| | AGE*YRSB | 3.24 |
| | AGE*WORK | 1.37 |
| | AGE | 1.80 |
| | WORK | 0.93 |
| | UGPA | 0.37 |
| | VERB*QUAN | 0.04 |
| | NOSC | 0.03 |
| | SCAV*QUAN | 0.02 |
| 20 | *QUAN | 19.80 |
| | *UGPA | 33.88 |
| | YRSB | 2.38 |
| | *VERB | 5.00 |
| | *SCAV | 4.15 |
| | YRSB*YRSB | 0.13 |
| | AGE*AGE | 0.71 |
| | WORK*WORK | 0.51 |
| | AGE | 0.53 |
| | WORK | 0.05 |
| | NOSC | 0.05 |
| 21 | *QUAN | 8.29 |
| | *UGPA | 35.30 |
| | *YRSB | 12.19 |
| | *VERB | 5.45 |
| | *QUAN*QUAN | 5.12 |
| | SCAV | 2.62 |
| | SCAV*SCAV | 2.25 |
| | AGE | 1.16 |
| | WORK | 1.03 |
| | NOSC | 0.06 |

Table 23. Regression with Six Best Variables, Male Sample.

| VARIABLE | MULTIPLE R | R SQUARE | B | BETA |
|---|---|---|---|---|
| QUAN | 0.27474 | 0.07548 | 0.079938 | 1.17992 |
| UGPA | 0.36092 | 0.13026 | 0.353521 | 0.31918 |
| YRSB | 0.45718 | 0.20901 | 0.032314 | 0.27007 |
| VERB | 0.47390 | 0.22458 | 0.008610 | 0.12831 |
| QUAN*QUAN | 0.48979 | 0.23990 | -0.001016 | -0.92797 |
| SCAV | 0.49979 | 0.24979 | 0.001216 | 0.10313 |
| (CONSTANT) | | | -0.631847 | |

| | |
|---|---|
| MULTIPLE R | 0.49979 |
| R SQUARE | 0.24979 |
| STANDARD ERROR | 0.34607 |

| ANALYSIS OF VARIANCE | DF | SUM OF SQUARES | MEAN SQUARE | F |
|---|---|---|---|---|
| REGRESSION | 6 | 10.76652 | 1.79442 | 14.98283 |
| RESIDUAL | 270 | 32.33658 | 0.11977 | |

Figure 70. Graph of standardized residuals versus $\hat{y}$, final regression, male sample.

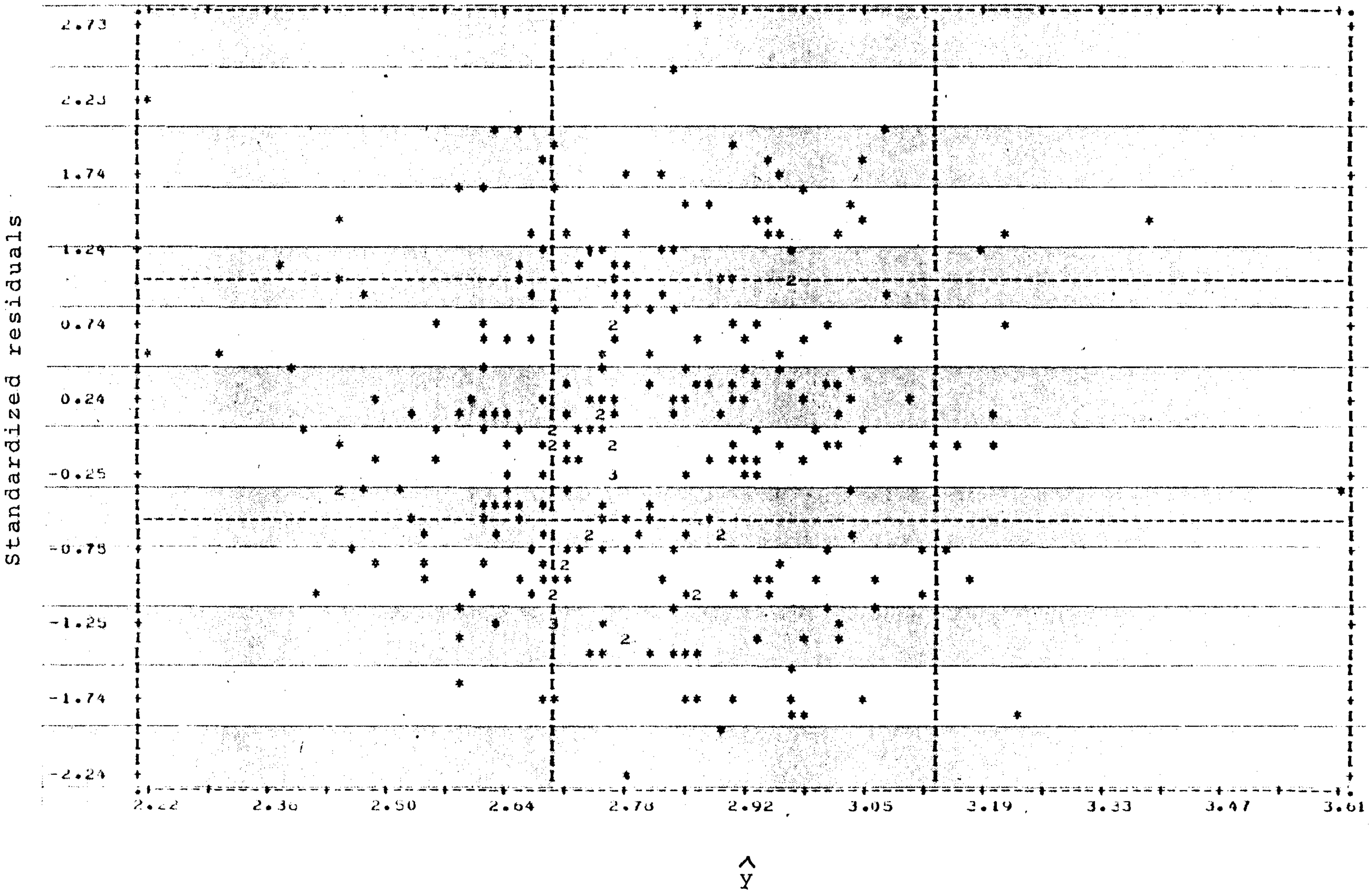


Figure 71. Graph of standardized residuals versus QUAN, final regression, male sample.

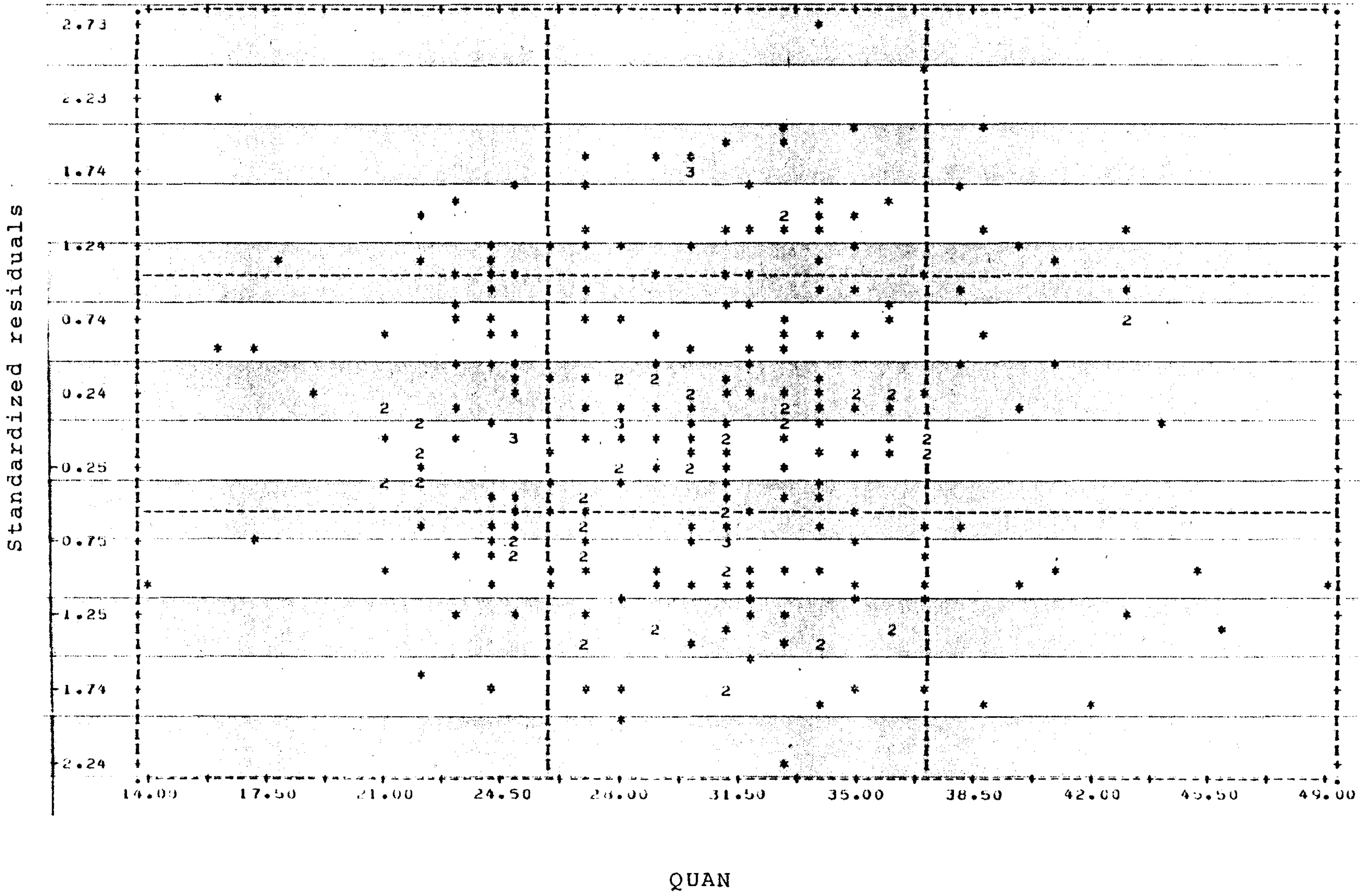


Figure 72. Graph of standardized residuals versus UGPA, final regression, male sample.

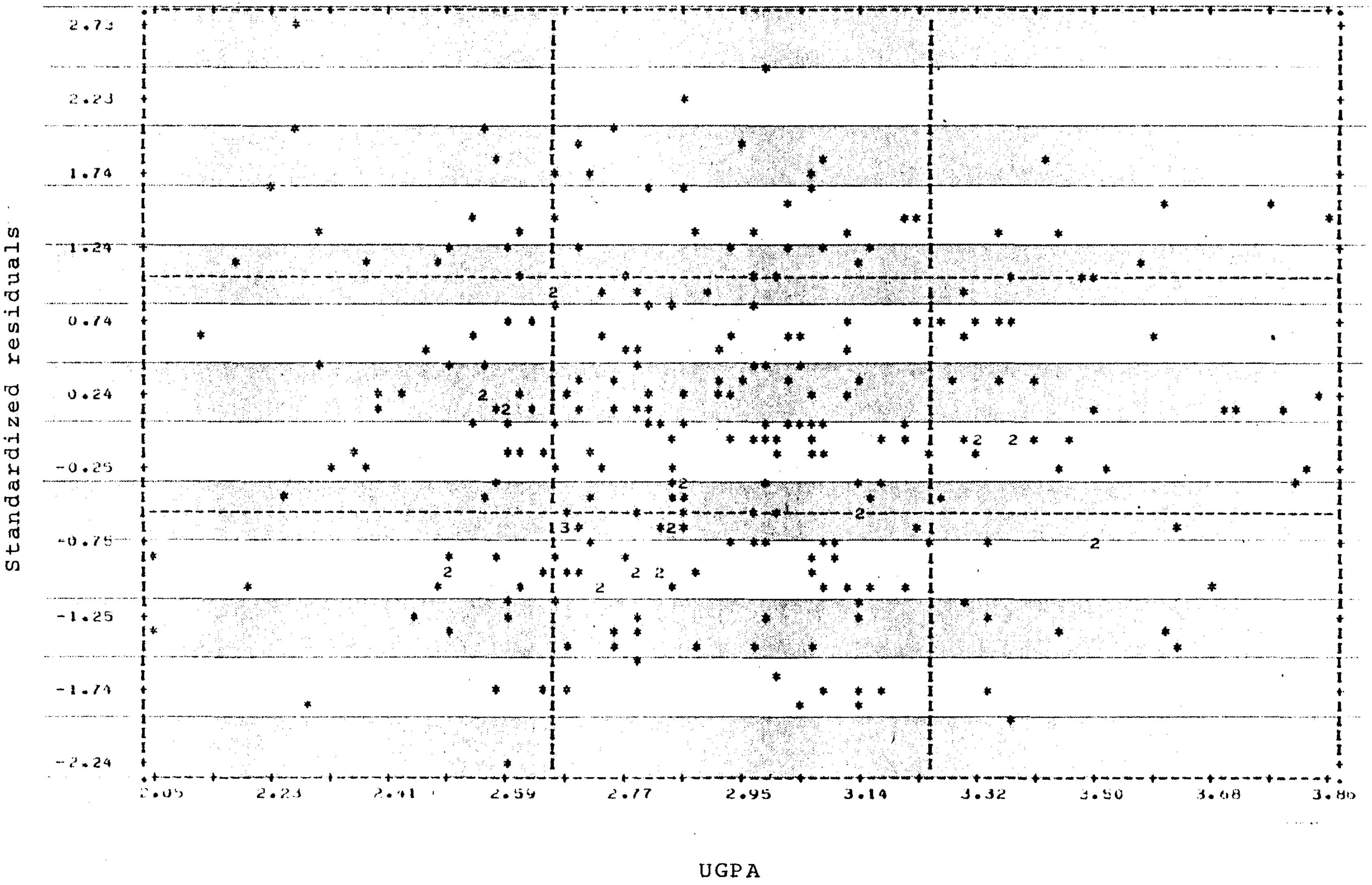


Figure 73. Graph of standardized residuals versus YRSB, final regression, male sample.

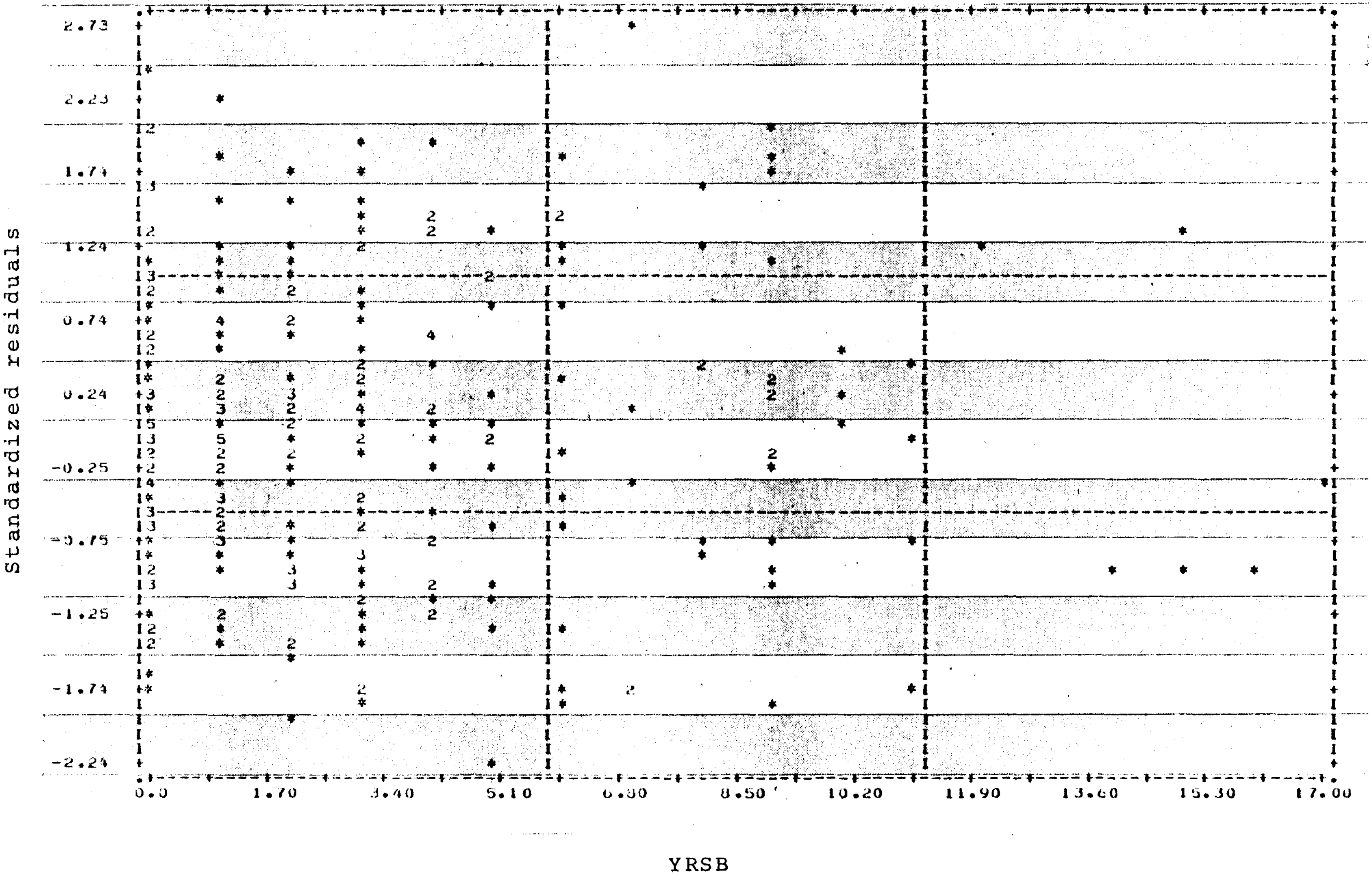


Figure 74. Graph of standardized residuals versus VERB, final regression, male sample.

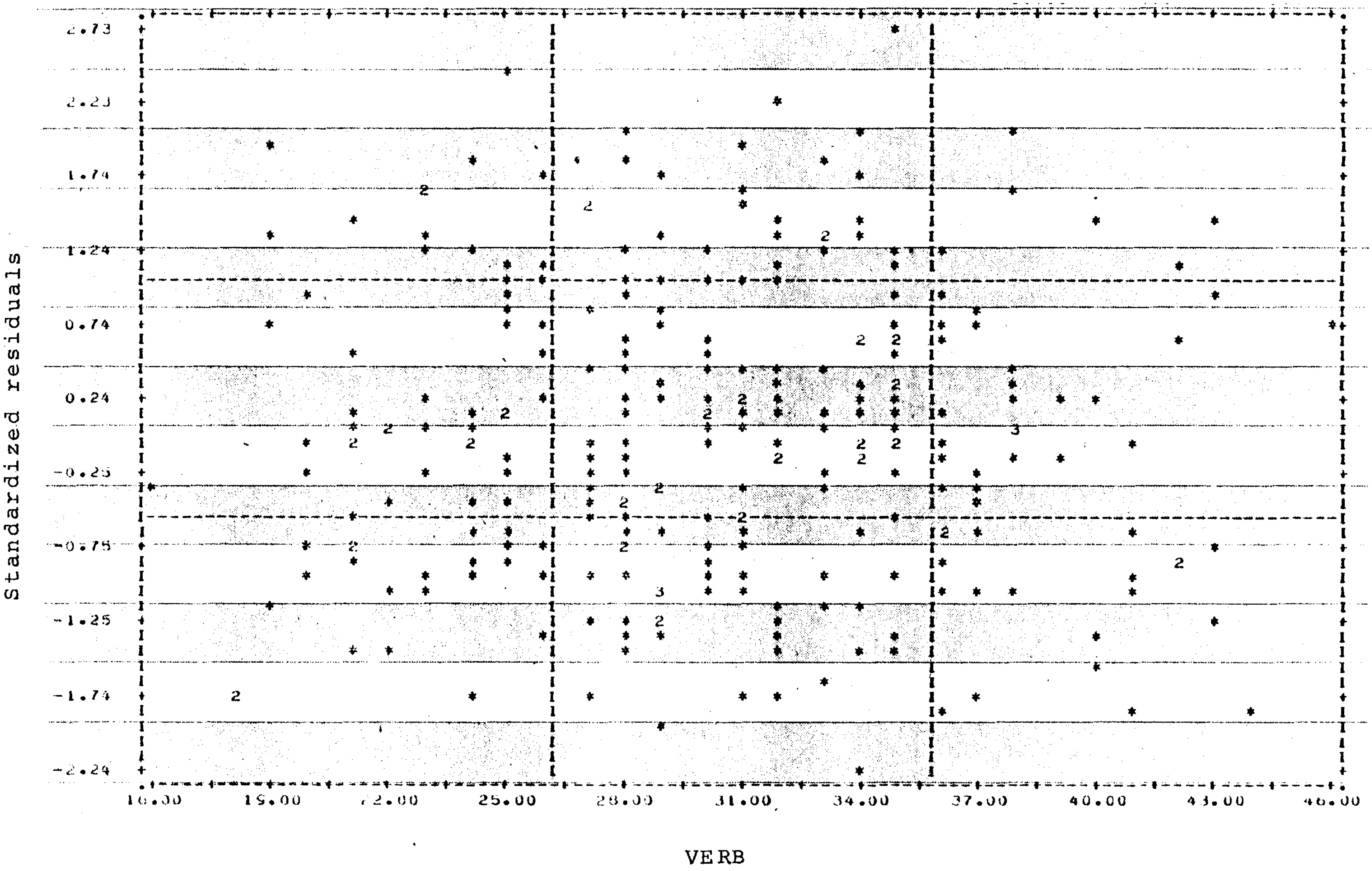


Figure 75. Graph of standardized residuals versus QUAN*QUAN, final regression, male sample.

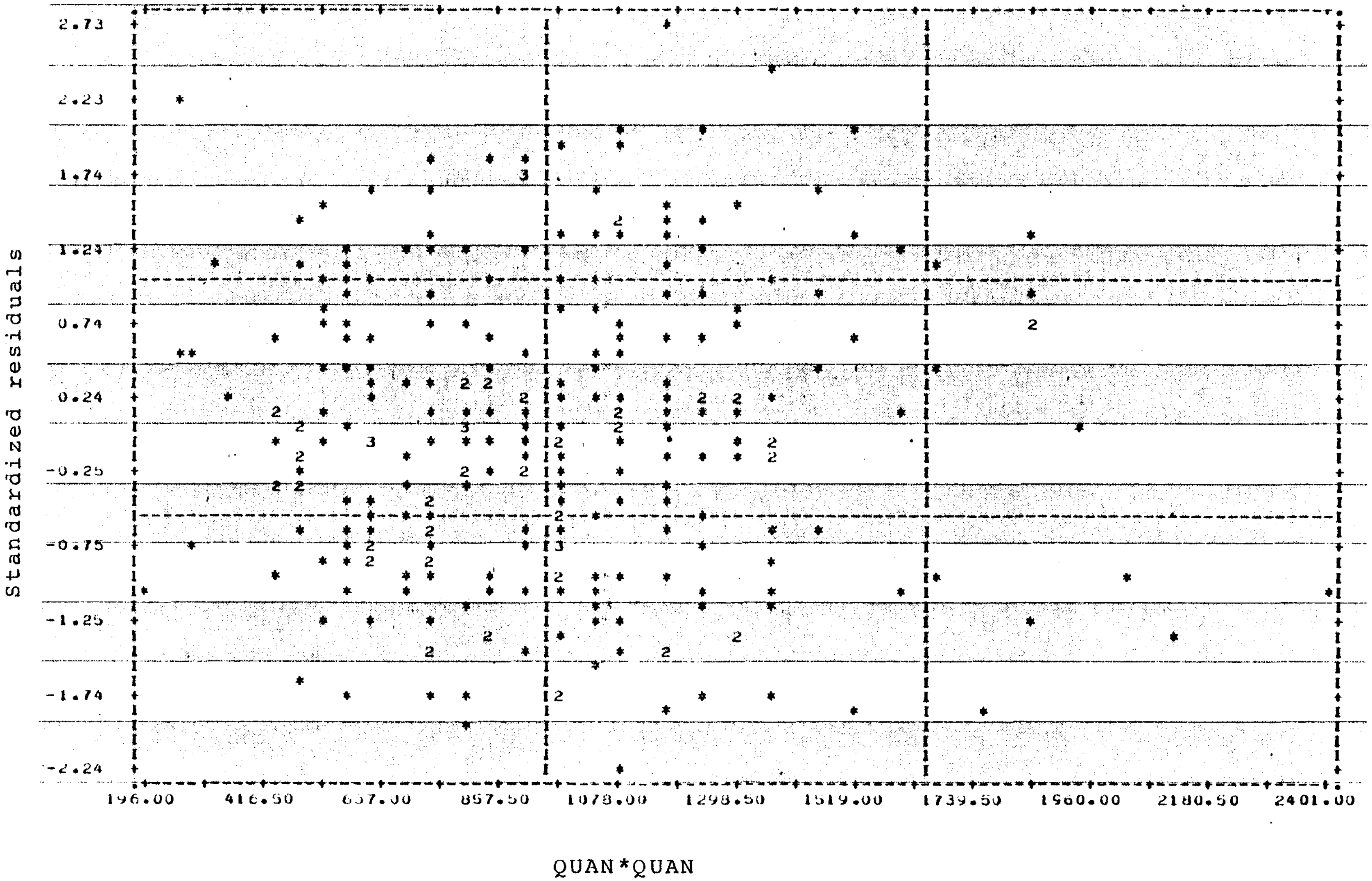
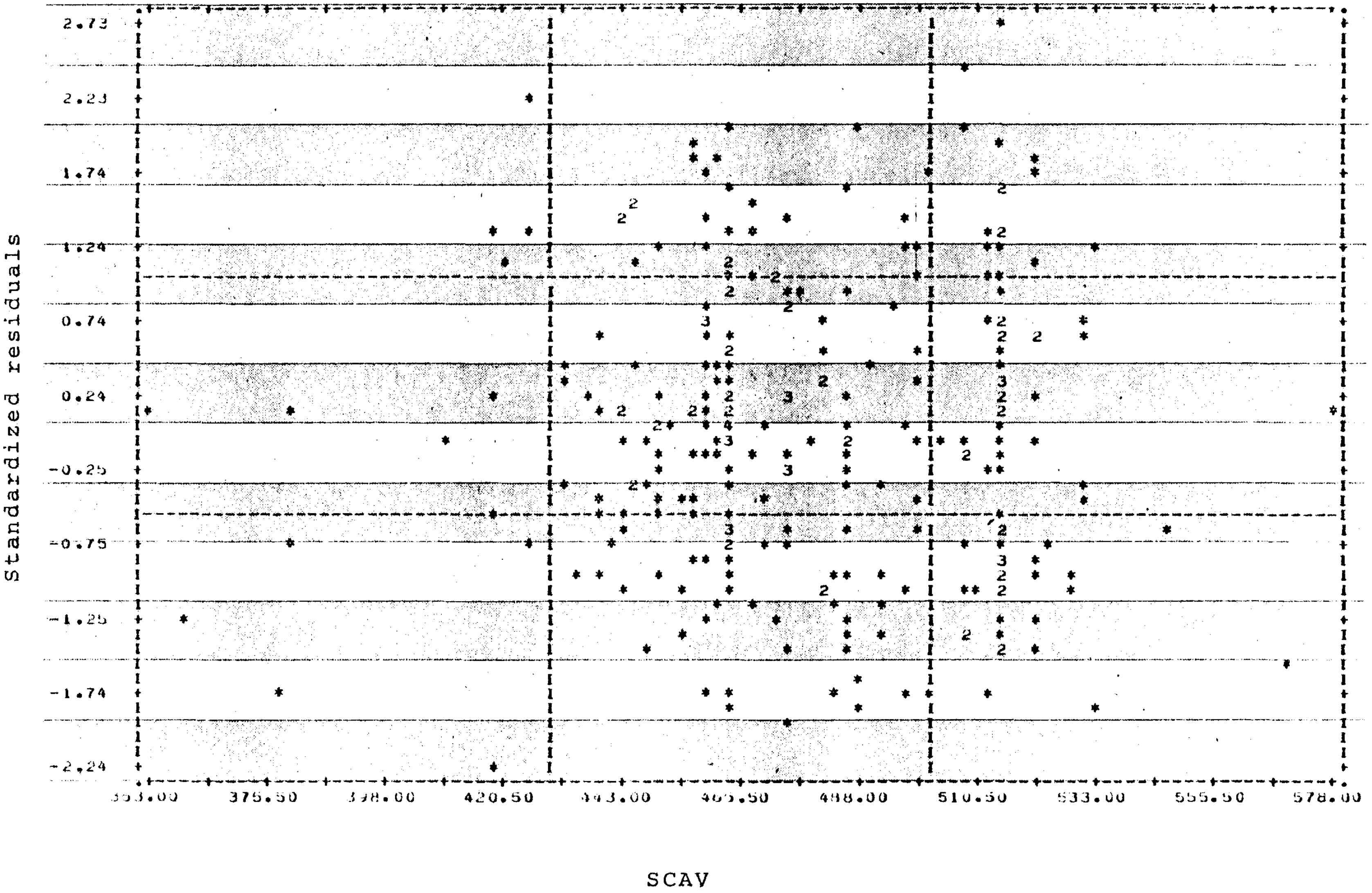


Figure 76. Graph of standardized residuals versus SCAV, final regression, male sample.



points have residuals greater in absolute value than two. In the plot of SCAV we see a slight underprediction in the first few graph points. In YRSB, this slight underprediction occurs in the last few points plotted.

The results from the male sample reinforce the results from previous studies. The squared multiple correlation is low, varying from 23% to 26%, not yielding much predictive accuracy. The scatterplots of residuals are randomly distributed about zero with most residuals falling between plus and minus two. The model is correct, but does not predict as well as would be desired.

# CHAPTER VI

# THE DISCUSSION

This chapter will consider first the results from the female sample. Then, the male sample's results will be analyzed. Finally, the two groups will be compared.

## The Female Sample, Variables

The initial regression on the female sample showed two problems. First, it gave low predictive accuracy. The value of $R^2$ was .13. Secondly, the residuals plots indicated that some model violations existed. Several non-typical points, outliers, were influencing the regression unfavorably. Removal of these points brought the value of $R^2$ up to .34 but the residuals plots again indicated that model violations remained. These violations appeared as a decrease in the scatter of the residuals as the variables AGE, WORK and YRSB increased. These patterns could have resulted from one of two causes: the need for additional polynomial variable terms, or the need for a transformation of the entire equation.

As a first try to improve the value of $R^2$, two dummy variables, DUMM and OTDG, were added to the regression. $R^2$ was increased only slightly, to .356, and the beta weights of both variables were small relative to those of the other variables.

Five interaction terms were tested on the next regression. $R^2$ increased, again only slightly, to .376. Of the five interaction terms, QUAN*UGPA, QUAN*SCAV, QUAN*VERB, AGE*WORK and AGE*YRSB, the ones having the greatest effect were QUAN*UGPA and AGE*WORK. These entered the equation first and had beta weights large in absolute value in comparison to the other weights in the equation.

The next regression used three squared terms, AGE*AGE, WORK*WORK and YRSB*YRSB, in addition to the original variables. These variables were the ones in which model violations had occurred in the plots of the standardized residuals. $R^2$ had a value of .379 for this program.

As a last program with added variables, those variables which had contributed most to the increase of $R^2$ in previous programs were combined with the original variables and a regression was run. These variables were OTDG, YRSB*YRSB, QUAN*UGPA and AGE*WORK. The value of $R^2$ was .389. So, with the removal of outliers, $R^2$ went from .13 to .34 and with terms added, $R^2$ went from .34 to .39. We had a total increase of $R^2$ from .13 to .39, an increase of .26.

The residuals plots of the last regression were examined and they showed that even though a better fit had been obtained through the elimination of outliers and the addition of other terms, still the model violations remained. The plots of AGE, YRSB and WORK still showed a decrease in scatter as the value of the independent variables increased. Since squared terms had not corrected this problem, it bacame evident that it was of a heteroscedastic nature and a transformation would need to be applied.

Since AGE, YRSB and WORK were all related concepts, it was hoped that a transformation for one of the variables would correct for the violations in all three variables. AGE was selected as the variable on which to base the transformation. The transformation used was 10/(36-AGE). Each term in the original regression equation was multiplied by this expression. A regression was run on the transformed equation. The resulting $R^2$ was .99. The residuals plots showed that the heteroscedasticity had been corrected for. The transformed variables and their F values and beta weights are shown in Table 24. With respect to beta weights, we see that the most influential terms were SCAV, QUAN and UGPA. YRSB was not significant either in F value or in beta weight. NOSC and VERB received negative weights. They acted to adjust for irrelevant variance, thus improving the fit. (10/(36-AGE)) was the transformation of the $\beta_0$ term and has an insignificant F value but the fourth largest beta weight in absolute value.

Table 24. F Values and Beta Weights, Transformed Equation.

| VARIABLE | F VALUE | BETA WEIGHT |
|---|---|---|
| 10SCAV/(36-AGE) | 32.128 | .647 |
| 10YRSB/(36-AGE) | 0.180 | .019 |
| 10QUAN/(36-AGE) | 73.128 | .326 |
| 10NOSC/(36-AGE) | 21.776 | -.081 |
| 10UGPA/(36-AGE) | 12.659 | .314 |
| 10WORK/(36-AGE) | 17.360 | .161 |
| 10VERB/(36-AGE) | 9.230 | -.142 |
| 10/(36-AGE) | 2.174 | -.321 |

Table 25. Comparison of Beta Weights for the Female Sample.

| VARIABLE | INITIAL REGRESSION | REGRESSION WITHOUT OUTLIERS | TRANSFORMED REGRESSION |
|---|---|---|---|
| QUAN | .205 | .370 | .327 |
| UGPA | .251 | .431 | .314 |
| YRSB | .140 | .139 | .019 |
| NOSC | -.112 | -.202 | -.081 |
| WORK | .362 | .350 | .161 |
| AGE | -.291 | -.046 | ---- |
| SCAV | .075 | .201 | .647 |
| VERB | .075 | -.051 | -.142 |

Table 26. F Values and Beta Weights, Last Male Regression.

| VARIABLE | F VALUE | BETA WEIGHT |
|---|---|---|
| QUAN | 9.202 | 1.180 |
| UGPA | 33.357 | .319 |
| YRSB | 24.368 | .270 |
| VERB | 5.554 | .128 |
| QUAN*QUAN | 5.700 | -.928 |
| SCAV | 3.558 | .103 |

## Comparison of First, Second and Last Regressions, Female Sample

A comparison of the beta weights for each term from the initial regression, the first regression with outliers removed and the transformed regression is shown in Table 25. In the initial regression, the variables with the largest positive beta weights were WORK, UGPA and QUAN. AGE had a large negative weight. SCAV and VERB had the smallest weights, approximately one-fifth the size of the beta weights given to WORK. The regression with outliers removed had three variables with comparatively large weights. These were UGPA, QUAN and WORK. The largest negative weight went to NOSC. In the transformed regression, the largest weights went to SCAV, QUAN and UGPA. WORK had the fourth largest beta weight. NOSC and VERB again had negative weights.

The most valuable variables for the female sample were QUAN, UGPA, SCAV and WORK. VERB and NOSC were suppressor variables. Though they had low correlations with the dependent variable, GGPA, their correlations with other variables caused them to act to eliminate irrelevant variance in the remaining variables. YRSB did not contribute much and could be eliminated. AGE, when entered, acted as a suppressor variable to improve the accuracy of the prediction.

The plots of the standardized residuals , in the initial regression, versus $\hat{y}$, UGPA, SCAV, QUAN, VERB and NOSC showed a random scatter in general with a few points which could be called outliers. The plots of the standardized residuals versus WORK, AGE and YRSB showed evidence of model violations, indicating that the multiple regression model could not be used as it was. In the regression with a transformation, the plots of the standardized residuals showed that these violations had been corrected. The majority of points in these graphs occurred in the first third or two-thirds of the plot, but neither these points not the remaining points gave any indication of a decrease in scatter indicative of heteroscedasticity. The multiple regression model was therefore appropriate in this situation. The analysis of the residuals plots enabled us to greatly improve the predictive accuracy of the multiple regression equation for females. The value of $R^2$ increased from .13 to .99 through the application of residuals analysis techniques. The plots of residuals first function was to locate outliers which were distorting the regression parameters. Secondly, they enabled us to spot model violations and gave us an indication of the direction to proceed in eliminating these violations and thus improving the fit of the model. The patterns suggested either a transformation or the addition of polynomial terms would correct the violations. The addition of terms did increase the value of $R^2$, but did not remove the heteroscedasticity. The transformation however, both improved $R^2$ and eliminated these model violations.

The Male Sample, Variables

The initial regression for the male sample used AGE, NOSC, YRSB, WORK, VERB, SCAV and UGPA as independent variables to predict GGPA. The value of $R^2$ was .239, a rather low value. The plots of the standardized residuals showed a few points which could be considered outliers, but otherwise showed no indication of model violations with regard to any of the variables. Three points were removed and the standardized residuals were again examined. The plots still exhibited a random scatter with most points falling between plus and minus two. In several of the plots, YRSB, SCAV and QUAN, a slight underprediction in one section of points indicated that squared terms of these variables might give a better fit. The main problem with the male sample, however, was simply a low predictive accuracy.

Several groups of terms were added to the initial group of independent variables and regressions were run. The first set of added terms included two dummy variables, DUMM and OTDG. The value of $R^2$ was .239 and neither new term added much to its value.

The second group of additional terms was a set of five interaction terms, QUAN*UGPA, QUAN*SCAV, QUAN*VERB, AGE*WORK and AGE*YRSB. The value of $R^2$ was .257, an increase of .018, but none of the F values of the interaction terms was significant (see Table 22).

Next, AGE*AGE, YRSB*YRSB and WORK*WORK were added. The value of $R^2$ was .242 and none of the variables had a significant F value.

The last new variables added to the original eight were QUAN*QUAN and SCAV*SCAV. The value of $R^2$ was .26 and QUAN*QUAN contributed significantly to this value. The plot of standardized residuals versus QUAN had originally suggested that the addition of QUAN*QUAN might improve the fit. None of the additions of new variables had increased $R^2$ by a great amount, so the variables were examined to see what maximum value of $R^2$ could be generated with the fewest number of variables. The F values of all the variables were examined (see Table 22) and those variables which had significant F values in any previous program were selected for inclusion in a final regression. These variables were QUAN, UGPA, YRSB, VERB, SCAV and QUAN*QUAN. The F value resulting from this equation was .25. So the best value with the fewest variables was .25, an increase of .011 over the initial value of $R^2$. This equation, however, used data collected on five variables rather than the eight originally used. The variables from this last regression are shown with their F values and beta weights in Table 26. Examiniation of this table shows that all variables but SCAV contributed significantly to the value of $R^2$.

The most valuable variables both in terms of F value and positive beta weight were QUAN, UGPA and YRSB. The beta weights of VERB and SCAV were low in comparison. QUAN*QUAN had a negative weight, indicating its function as a suppressor variable.

The plots of standardized residuals for this program showed no model violations in the selection of variables. The points are randomly scattered with all but four falling between plus and minus two.

## Comparison of Initial and Final Regressions, Male Sample

The initial regression for males used data collected on seven variables, the final regression used data collected on five variables, a sixth term was formed by squaring one variable. The most important variables, both in terms of F value and beta weight were QUAN, UGPA and YRSB, in both programs. The importance of QUAN seems to reflect the relevance of mathematical training to the business program. The MBA degree is one which requires and uses calculus, statistics and other mathematical areas in many of its courses. YRSB was significant whereas neither AGE nor WORK was; perhaps this reflects some maturity factor that one gains from staying out of school for a while. The importance of UGPA would imply a consistency of achievement across class levels, a good undergraduate performance implying a good graduate performance.

Each program had one negative weight variable, AGE in the initial one and QUAN*QUAN in the final one. In both programs, VERB received a low beta weight in comparison to other terms, though its F value was significant in both cases. SCAV was not only given a low beta weight, but an insignificant F value as well, in both programs. The calculated parameters of variables common to both programs are very similar. The value of $R^2$ increased from .239 to .250 and the standard error decreased slightly from .3495 to .3461. The final value of $R^2$ was not much greater than the initial value. This finding tends to substantiate those of previous research. The multiple regression model is an appropriate one, judging from the plots of the standardized residuals, but does not yield much predictive accuracy for the male sample.

## Comparison of the Female and Male Samples

The initial regressions for males and females both showed low values of $R^2$, .239 and .135, respectively. The values of the parameters and beta weights generated by each regression are shown in Table 27. We see that all parameters vary from sample to sample. The largest differences, as seen in the standardized weights, occur in YRSB, VERB and WORK. Moderate differences occur in AGE,

Table 27. Comparison of B Values and Beta Weights, Both Samples.

B VALUES

| VARIABLE | MALES | FEMALES | DIFFERENCE |
|---|---|---|---|
| QUAN | .0167 | .0154 | .0013 |
| UGPA | .3449 | .2758 | .0691 |
| YRSB | .0349 | .0135 | .0214 |
| VERB | .0087 | -.0055 | .0142 |
| SCAV | .0012 | .0008 | .0004 |
| AGE | -.0177 | -.0247 | .0070 |
| WORK | .0015 | .0037 | -.0022 |
| NOSC | ----- | -.0454 | ----- |

BETA WEIGHTS

| VARIABLE | MALES | FEMALES | DIFFERENCE |
|---|---|---|---|
| QUAN | .2546 | .2045 | .0501 |
| UGPA | .3137 | .2505 | .0632 |
| YRSB | .3018 | .1398 | .1620 |
| VERB | .1307 | -.0748 | .2055 |
| SCAV | .1074 | .0751 | .0323 |
| AGE | -.1881 | -.2912 | .1031 |
| WORK | .1408 | .3623 | -.2215 |
| NOSC | ----- | -.1125 | ----- |

QUAN and UGPA. The smallest difference was in SCAV. NOSC cannot be compared since it did not enter the male equation as it did in the one for females.

A comparison of the scatterplots of the standardized residuals showed that the male group had fewer outliers, three out of two hundred eighty, as compared to eleven out of one hundred twenty for the female group. Also, the male scatterplots showed no evidence of model inappropriateness while the female sample indicated heteroscedasticity in three of the variables. The multiple regression model, therefore, could not be used for the female sample without adjustment.

The regression results for the female sample showed it was very responsive to improvement techniques. Removal of outliers increased the value of $R^2$ from .135 to .342. Additional terms brought $R^2$ up to .389. A transformation increased $R^2$ further up to .99. The male sample, on the other hand, showed little response and supported prior research findings. The value of $R^2$ increased only from .23 initially to .250 in the final regression.

In the final regressions for each sample, we see even greater differences than initially. The program for males had a predictive accuracy of 25%, for females this accuracy was 99%. The equation for females required a transformation whereas that for males used five variables as they originally occurred plus one squared term. The three most important variables for predicting in the male sample were QUAN, UGPA and YRSB; in the female sample they were QUAN, UGPA and SCAV. YRSB was not significant in the female sample. SCAV had a low weight in the male sample. VERB had a low weight in the equation for males and a negative weight in the equation for females. Overall, VERB was not nearly as significant a variable as QUAN for either sample.

If the male equation were used to predict female performance, then, for the student with variable values shown on page 107, the equation for males would predict a value of 3.43 while the equation for females predicted 2.74. Because the number of females enrolled in MBA programs is traditionally much smaller than the number of males, if they were incorporated into the male sample and only one equation were generated, the male input would dominate and the results could be highly misleading for the female group.

# SUMMARY

Multiple regression equations are frequently used in the prediction of academic performance. However, the graphs of the standardized residuals generated from the regressions have not been examined to see whether or not the model has been appropriately specified and no violations exist. This study investigated the appropriateness of the multiple regression model for prediction of student grade point averages in the Graduate School of Business through analysis of the graphs of standardized residuals versus the fitted and independent variables. The model was also checked to see whether the predictive accuracy could be increased either by application of techniques suggested by the graphs of the standardized residuals or by the addition of variables based on information in students' files.

The students were separated into two groups on the basis of sex and each group was analyzed separately. The multiple regression model was found to be correctly specified for the male sample, but for the female sample, violations existed which made the model inappropriate. Through the use of a transformation (the method of weighted least squares), the existing heteroscedasticity was removed and the adjusted model was then appropriate

for the female sample.

Two variables, the quantitative score on the GMAT and the undergraduate grade point average, were important for both groups. The number of years since the undergraduate degree was awarded was a major variable for the male group, but insignificant for females. The school quality index had a high weight for females, but a low one for males.

The final value of the multiple correlation coefficient squared for the male sample was .25; for the female sample it was .99. The analysis of the plots of standardized residuals and the separation of the sample by sex enabled us to generate a multiple regression equation for females with high predictive accuracy and no evidence of model violations. The results for the male sample, however, remained low though the residuals analysis showed the model had been appropriately applied.

# REFERENCES

Anastasi, A. Psychological testing (4th ed.). New York: Macmillan, 1976.

Anscombe, F. J. Graphs is statistical analysis. American Statistician, 1973, 27, 17-21.

Astin, A. A. Predicting academic performance in college. New York: The Free Press, 1971.

Beckham, T. W. A differential weighting of the undergraduate grade point average as a method of improving the procedure for selecting students for dental school. Unpublished doctoral dissertation, Loyola University of Chicago, 1973.

Boldt, R. F. Discrepant predictor study. Princeton, N. J.: Educational Testing Service, Brief Number 2, 1969.

Burnham, P. S. & Hewitt, B. A. Secondary school grades and other data as predictors of academic achievement in college. College and University. 1972, 48, 21-22.

The Carnegie Council on Policy Studies in Higher Education. Fair practices in higher education: rights and responsibilities of students and their colleges in a period of intensified competition for enrollments. San Francisco: Jossey-Bass, 1979.

Casserly, P. L. & Campbell, J. T. A survey of skills and abilities needed for graduate study in business. Princeton, N. J.: Educational Testing Service, Brief Number 9, 1973.

Chatterjee, S. & Price, B. Regression analysis by example. New York: John Wiley & Sons, 1977.

Connelly, F. J. & Nord, W. R. A study of the influence of undergraduate course content on the admission test for graduate study in business as a predictor of success in graduate schools of business. Princeton, N. J.: Educational Testing Service, Brief Number 5, 1972.

Cooley, W. W. Techniques for considering multiple measurement. In R. L. Thorndike (Ed.), Educational measurement (2nd ed.). Washington, D. C.: American Council on Education, 1971.

Daniel, C. & Wood, F. S. Fitting equations to data. New York: Wiley-Interscience, 1971.

Darlington, R. B. Multiple regression in psychological research and practice. Psychological Bulletin, 1972, 69, 161-182.

Dawes, R. M. A case study of graduate admissions: application of three principles of human decision making. American Psychologist, 1971, 26(2), 180-188.

Dawes, R. M. Graduate admission variables and future success. Science, 1975, 187(4178), 721-723.

Draper, N. R. & Smith, H. Applied regression analysis. New York: John Wiley & Sons, 1966.

Ebel, R. L. Essentials of educational measurement. Endlewood Cliffs, N. J.: Prentice-Hall, 1972.

Educational Testing Service. GMAT statistical summary by undergraduate colleges attended, 1957-1976. Princeton, N. J.: Educational Testing Service, 1979.

Engelhart, M. D. Methods of educational research. Chicago: Rand McNally, 1972.

Finn, J. D. A general model for multivariate analysis. Chicago: Holt, Rinehart and Winston, 1974.

Fishman, J. A. & Pasanella, A. K. College admission-selection studies. Review of Educational Research, 1960, 30, 298-310.

Gadzella, B. M., Cochran, S. W., Parham, L. & Fournew, G. P. Accuracy and differences among students in their predictions of semester achievement. Journal of Educational Research, 1976, 70(2), 75-81.

Givner, N. & Hynes, K. Admissions test validity: correcting for restriction effects. College and University, 1979, 54(2), 119-123.

Hays, W. L. Statistics for the social sciences (2nd ed.). New York: Holt, Rinehart and Winston, 1973.

Kendall, M. G. & Stuart, A. The advanced theory of statistics (Vol. 3). London: Charles Griffin, 1968.

Kim, J. & Kohout, F. J. Multiple regression analysis: subprogram regression. In Nie, J. J., Hull, C. H., Jenkins, J. G., Steinbrenner, L. & Bent, D. H. Statistical package for the social sciences (2nd ed.). New York: McGraw-Hill, 1975.

Kim, J. & Kohout, F. J. Special topics in general linear models. In Nie, N. J., Hull, C. H., Jenkins, J. G., Steinbrenner, L. & Bent, D. H. Statistical package for the social sciences (2nd ed.). New York: McGraw-Hill, 1975.

Lavin, D. E. The prediction of academic performance: a theoretical analysis and review of research. New York: Russell Sage Foundation, 1965.

Maxwell, S. E. & Jones, L. V. Female and male admission to graduate school: an illustrative inquiry. Journal of Educational Statistics, 1976, 1, 1-37.

Misanchuk, E. R. A model-based prediction of scholastic achievement. Journal of Educational Research, 1977, 71(1), 30-35.

Mitchell, J. L. Developing new role relationships for men and women in business. Journal of Community and Organizational Development, 1979, 1(2), 9-11.

Mosteller, F. & Tukey, J. W. Data analysis and regression. Reading, Massachusetts: Addison-Wesley, 1977.

Nunnally, J. C. Psychometric theory. New York: McGraw-Hill, 1967.

Petry, J. R. & Craft, P. A. Investigation of instruments to predict academic performance of high-risk college students. Journal of Educational Research, 1976, 70(1), 21-25.

Pitcher, B. Predicting first-year average grades of interrupted and uninterrupted students in graduate business schools. Princeton, N. J.: Educational Testing Service, Brief Number 9, 1973.

Pitcher, B. & Schrader, W. B. Indicators of college quality as predictors of success in graduate schools of business. Princeton, N. J.: Educational Testing Service, Brief Number 6, 1972.

Pitcher, B. & Smith, H. Moderator variable study: the effect of background factors on the prediction of performance in graduate business school. Princeton, N. J.: Educational Testing Service, Brief Number 3, 1969.

Powers, D. E. & Evans, F. R. Relationships of preadmission measures to academic success in graduate management education. (ETS RB-78-11). Princeton, M. J.: Educational Testing Service, 1978.

Rao, C. R. Linear statistical inference and its applications (2nd ed.). New York: John Wiley & Sons, 1973.

Schrader, W. B. The graduate management admission test: technical report on test development and score interpretation for GMAT users. Princeton, N. J.: Educational Testing Service, 1979.

Schwartz, M. J. & Clark, F. E. Prediction of success in graduate school at rutgers university. Journal of Educational Research, 1959, 53, 109-111.

Searle, S. R. Linear models. New York: John Wiley & Sons, 1971.

Tukey, J. W. Exploratory data analysis. Reading, Massachusetts: Addison-Wesley, 1977.

Weinstein, E. L., Brown, I. & Wahlstrom, M. W. A systems view of admissions procedures. College and University, 1979, 54(2), 124-138.

Wesolowsky, G. O. Multiple regression and analysis of variance. New York: John Wiley & Sons, 1976.

Wiggins, J. A. Hypotheses validity and experimental laboratory methods. In H. M. Blalock, Jr. & A. B. Blalock (Eds.), Methodology in social research. New York: McGraw-Hill, 1968.

Wilson, K. V. Linear regression equations as behavior models. In J. R. Royce (Ed.), Multivariate analysis and psychological theory. New York: Academic Press, 1973.

Wikoff, R. L. & Kafka, G. F. Interrelationships between the choice of college major, the ACT and the sixteen personality factor questionnaire. Journal of Educational Research, 1978, 71(6), 320-324.

# APPROVAL SHEET

The dissertation submitted by Mary E. Malliaris has been read and approved by the following committee:

Dr. Jack A. Kavanagh, Director
Associate Professor, Education, Loyola

Dr. John M. Wozniak
Professor, Education, Loyola

Dr. Samuel T. Mayo
Professor, Education, Loyola

Dr. Ronald R. Morgan
Associate Professor, Education, Loyola

The final copies have been examined by the director of the dissertation and the signature which appears below verifies the fact that any necessary changes have been incorporated and that the dissertation is now given final approval by the Committee with reference to content and form.

The dissertation is therefore accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

4/20/81
Date

Jack A. Kavanagh
Director's Signature