Remodeling Load Time Variability Data

Master's Writing Project

by

Michael J. Amdahl

August 1993

The author thanks Dr. John Borkowski for his help in making *this* writing project possible.

INTRODUCTION

The Postal Rate Commission (PRC) is, in part, concerned with modeling data derived from various segments of the United States Postal Service (USPS). Certain attributes of the model (e.g.the regression coefficients) are used in other functional forms to establish a cost analysis. The cost analysis is used by the PRC to help determine postal rates. From one of the USPS segments, data pertaining to Load Time Variability (LTV) has been collected on each of three types of delivery sites: Single Delivery Residential (SDR), Business and Mixed (BAM), and Multiple Delivery Residental (MDR). Load Time is defined as: "The time spent at delivery sites actually making deliveries, including incidental time for customer contacts, special services, and collections from delivery boxes. For cost analysis, load time is regarded as two distinct components: elemental load time and coverage-related load time."(PRC Doc) Elemental load time varies directly with the number of pieces of mail delivered to delivery points, and can be thought of as the sensitivity of load time to increases in volume at a stop that is already accessed or "covered". It is desired to derive a sensible regression model for this component of Load Time. Traditionally, this was a model that possessed a large adjusted r-squared statistic (R_a^2).

In this study, R_a^2 is obtained through a linear regression analysis of data recorded in the LTV study, and parameter estimation is an important consideration in building the regression model. After the parameter estimates have been found, they will be used by the PRC to calculate elasticities. Elasticities are functions of the model, and the coefficients derived from the model, and are sensitive to changes in the coefficients, this implies that the individual contribution of each regression coefficient is important. It is important to note that these data have been previously analyzed by four independent parties external to the PRC, referred to as "witnesses", and each witness proposed a model to the PRC. A considerable effort has been spent in an attempt to justify which model should be used, and subsequently, which coefficients should be retained in that model. The model's R_a^2 statistic is the basic criterion for a model being considered "good". Hence, one model is better than another if it possesses a larger R_a^2 than a competing model. Model assumptions and desirable characteristics associated with a "good" regression model have been ignored by the PRC witnesses. It is quite possible a better model exists if certain diagnostics checking model assumptions, and other important considerations are taken into account.

REGRESSION MODELS: NOTATION AND THEORY

A regression model can be expressed, in matrix notation, as

$$\mathbf{Y} = \mathbf{X}\beta.$$

$\mathbf{Y}$ is an $n \times 1$ vector of responses (dependent variables), $\mathbf{X}$ is an $n \times p$ matrix of covariates (independent variables), β is a $p \times 1$ vector of regression coefficients (parameters); and ϵ is an $n \times 1$ vector of random error terms which accounts for unexplained variation in the model. It is assumed $\epsilon_i \sim (0,\sigma^2)$, where the value of σ^2 is unknown. The expected value of $\mathbf{Y}$ (E[$\mathbf{Y}$]) is equal to $\mathbf{X}\beta$. If y_i is independent of y_j (for i $\neq$ j), then the covariance between y_i and y_j (Cov(y_i, y_j)) is equal to zero. This implies the Var($\mathbf{Y}$) = Var(ϵ) = σ^2I. The goal of this regression analysis is to obtain an estimate for the unknown parameter vector β. The regression coefficients (the individual components of β) will be estimated by a method called least squares regression of $\mathbf{Y}$ on $\mathbf{X}$.

In the method of least squares regression, $\mathbf{b}$ denotes the estimate of β, i.e., the $p \times 1$ vector of estimated coefficients of $\mathbf{X}$ that is derived as the least squares solution to the normal equations. That is, $\mathbf{b} = (\mathbf{X'X})^{-1}\mathbf{X'Y}$ is the solution to the normal equations: $\mathbf{X'Xb} = \mathbf{X'Y}$. In this setting, $\mathbf{X}$ is chosen to have full column rank which implies the columns of $\mathbf{X}$ are linearly independent (implying the existence of $(\mathbf{X'X})^{-1}$). Since $(\mathbf{X'X})^{-1}$ exists, $\mathbf{b}$ is the *unique* solution to the normal equations. Once $\mathbf{b}$ has been determined, the fitted model can be written as: $\hat{Y} = \mathbf{Xb}$. It is known that $\mathrm{E}[\hat{Y}] = \mathrm{E}[\mathrm{Xb}] = \mathbf{X}\beta$, implying that $\hat{Y}$ is an unbiased estimator of the expected value of $\mathbf{Y}$. Also, $\mathrm{Var}(\hat{\mathbf{Y}}) = \mathrm{Var}(\mathbf{Xb}) = \sigma^2\mathbf{X}(\mathbf{X'X})^{-1}\mathbf{X'}$. The distribution of $\hat{\mathbf{Y}}$ can be compactly expressed as $\hat{\mathbf{Y}} \sim (\mathbf{Xb}, \sigma^2\mathbf{X}(\mathbf{X'X})^{-1}\mathbf{X'})$. After the fitted model has been obtained, by the method of least squares, a measure of goodness of fit can be determined.

The R_a^2 statistic is one measure of goodness of fit of a regression model to the data. The mathematical formula for R_a^2 can be expressed as

$$R_a^2 = 1 - \frac{(n-1)SSE}{(n-p)SSR} \quad p < n$$

where n = number of observations, p = the number of regression coefficients, and SSE and SSR are the sums of squares for error and regression which can be obtained through a partitioning of SST. Where $SST = \mathbf{Y'Y}$. SST is partitioned into two pieces, one belonging to the column space of $\mathbf{X}$ and the other belonging to the null space of $\mathbf{X'}$. This can be written as

$$\mathbf{Y'Y} = \mathbf{Y'HY} + \mathbf{Y'(I-H)Y}.$$

$\mathbf{H}$ is a symmetric and idempotent matrix ($\mathbf{H} = \mathbf{H}' = \mathbf{H}^2$) and is called a perpendicular projection operator (ppo). $\mathbf{H}$ projects a vector (in this case $\mathbf{Y}$) onto the column space of $\mathbf{X}$ along the null space of $\mathbf{X}'$. $\mathbf{H}$ can be expressed as

$$\mathbf{H}= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'.$$

SSR, the sum of squares due to regression, is equal to $\mathbf{Y'HY}$. SSE, the sum of squares due to error, is equal to $\mathbf{Y'(I-H)Y}$. A perfect fit of the model to the data results in $R_a^2 = 1$, and implies the fitted model does not contain any error (SSE=0). $R_a^2 = 1$ is an idealized value (not one that is likely to be observed in practice). The closer R_a^2 is to one, the better the fit. However, no matter how good the fit is, certain assumptions are made when a regression model is fit to the data.

MODEL ASSUMPTIONS AND DIAGNOSTICS

Two basic assumptions in *this* regression analysis are: 1) the data have been recorded without error (at least negligible error), and 2) the n components of ϵ are independent and identically distributed (iid) as normal random variables with mean zero (μ=0) and constant variance ($\sigma_i^2 = \sigma_j^2$ for all i,j). These data have been recorded by someone other than the person doing the regression analysis. Therefore, the person doing the regression analysis should question the validity of the first assumption. A non-modeling error occurs if the first assumption has been violated, and may lead to the fitting of a useless model. The model could be useless in that it would not be explaining a phenomenon the analyst thought it

-$h_i\sigma^2$, and Var(e_i) depends on the location of $\mathbf{X_i}$ in the regressor space.

The studentized residuals are more helpful in checking the second assumption since they behave more like the ϵ_i. Denote the ith studentized residual by t_i. Then,

$$t_i = \frac{e_i}{s(1-h_i)^{1/2}}$$

where s denotes the root mean square error ($\sqrt{MSE}$). An approximation for the expected value of ith order statistic is given by

$$z_i = \frac{(i-0.375)}{(n+0.25)}$$

.

A Quantile-Quantile plot (Q-Q plot) of t_i vs. the expected value of the ith order statistics should show a straight line with a slope of approximately one, and would be evidence that the normality assumption has not been violated. A plot of t_i vs $\hat{y}_i$ should show a random scatter of points about the zero horizontal line. As a note, there are numerous methods (definitely more rigorous) available for checking model assumptions, but the above mentioned methods were used for their simplicity. Besides checking model assumptions, it is important to address the problem of multicollinearity.

Multicollinearity is an important consideration since the individual contribution of each regressor variable is of interest. In the presence of multicollinearity, the individual effect of a regressor variable can be masked by the effects of one or more of the other regressor variables in the model. The regressor variables can be thought of as independent variables, was explaining or wanted it to help explain. I suggest (prior to fitting the model) that some type of data checking be conducted so that one may feel reasonably confident that the first assumption has not been violated. Preliminary plotting of the data (such as histograms of the regression variables) may call attention to unusual or suspect observations. If some data points appear to be recorded in error, the analyst should bring this to the attention of the experts in the field. Data points should not be excluded from the data set unless there is a good reason for doing so. Extreme observations are not necessarily the result of errors in recording the data, they may simply represent rare observations. For example, suppose Z $\sim$N(0,1), then P(Z$\geq$ 2.0)=0.0275. If z =2.0 is observed, this does not imply that z =2.0 is an error, this simply means it is an unusual observation. However, if z =20 is observed we might be more inclined to think this value of Z is an error since the probability of observing a value this extreme or more extreme is zero to 88 decimal places. In any case, the existence of extreme observations should prompt further investigation into the nature of the anomaly. In order to check the validity of the second assumption, the fitted model is required.

Again, the fitted model is $\hat{Y}$ =**Xb**. In the regression setting, the residual vector ($\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}}$) is defined as the difference between the observed **Y** and the fitted **Y**. It can be written as $\mathbf{e}$ = **(I-H)Y**, and $\mathbf{e}\sim$ SN(0,(I-H)σ^2). That is, **e** has a singular normal distribution. This is different than the distribution of ϵ (which is normally distributed). The ith residual is defined by $e_i = y_i - \hat{y}_i$, and $e_i \sim$ SN(0,(1-h_i)σ^2), h_i denotes the ith diagonal element of **H** and $0 \leq h_i \leq 1$. Although we can not observe ϵ we can observe the residual vector (**e**), and will use it to check the second assumption. However, the e_i's are not iid since Cov($e_i e_j$) =

but in a certain sense there may be linear dependencies among the regressor variables. When the regressor variables are not truly independent of each other, then collinearity (or multicollinearity when more than two variables are involved) is said to exist among the regressor variables. The problem of multicollinearity among the regressor variables can be addressed by inspection of the Variance Inflation Factor (VIF) associated with each term. The VIF of the ith variable ($(VIF)_i$) is related to the correlation that exists between that ith variable and the remaining variables in the model.

$$(VIF)_i = \frac{1}{(1 - r_i^2)} \qquad 1 \leq (VIF)_i < \infty$$

where r_i^2 denotes the squared multiple correlation (coefficient of multiple determination) between the ith regressor variable and all other regressor variables. The optimal size for a VIF is one. A VIF equal to one for the ith predictor implies the ith regressor variable is uncorellated with the remaining regressor variables in the model. As a rule of thumb, a variable with a VIF over 10 is considered bad, and should be dropped from the model or an alternative to the method of least squares should be considered. A VIF greater than 10 implies the coefficient of multiple determination of the regression produced by regressing the ith regressor variable againest the other regressor variables exceeds 90 percent. A large VIF for the regressor variable x_i implies the accuracy of the estimate b_i is compromised. Since the goal of building these regression models is to use them for parameter estimation, it is very important to diagnose multicollinearity between the regressor variables.

DIAGNOSTICS ON PREVIOUSLY PROPOSED MODELS

I will perform diagnostics on the models selected as the best linear models proposed to the PRC by the PRC witnesses. As a note, the response variable for each model is DTSUM.

For the BAM model, 20 regressor variables have been used, and R_a^2 is reported as 0.814. Diagnostics of the model show that 8 variables have VIF's over 10; PDS2 and LDPDS have VIF's over 11000. Since a VIF of 10000 for the ith variable implies r_i^2 is .9999, there is an extremely serious collinearity problem associated with these two variables. In all, 40 percent of the estimated coefficients are unreliable as judged by their VIF's. CT1 has a p-value of .0759, and FVDC has a p-value of .0839 this implies that these two variables are not very helpful as explanatory variables. The variable FVDC is not significant and has a VIF is over 10. Page 21 has a list of the model variables, VIF's and p-values. The Q-Q plot given on page 22 exhibits an s-shape suggesting that the normality assumption is questionable.

For the SDR model, 26 regressor variables have been used, and R_a^2 is reported as 0.352. 6 terms have VIF's over 10. The largest VIF is 159.2 (associated with AVDC), and the smallest VIF is 10.4 (associated with VC2). Hence, collinearity is a problem in 23 percent of the regression variables. The two largest p-values are reported as .0623 from variable FVDC (associated VIF of 19.5), and .0139 from variable VCPDS (associated VIF of 8.4). 22 of the variables have very small p-values (reported as .0001). Page 25 has a list of the model variables, VIF's and p-values. Although the collinearity problem associated with the SDR data set is the least severe among the models, the R_a^2 statistic is the smallest in magnitude. An R_a^2 this small implies the SDR model does the worst job of explaining the variation in $\mathbf{Y}$. The Q-Q plot, presented on page 26, shows the presence of a heavily-tailed distribution.

Hence, the normality assumption is questionable.

For the MDR model, 24 regressor variables have been used, and R_a^2 is reported as 0.916. Page 23 has a list of the model variables, VIF's and p-values. 16 of the variables have VIF's over 10. The largest VIF is 97.3 (from LADD). Collinearity is a problem with 66% of the variables, this implies the estimated coefficients are extremely unreliable. The two largest p-values are 0.0953 (from LADD) and 0.0844 (from VCPDS). Again, we find a variable (LADD) that is insignificant and has a large VIF. The Q-Q plot (page 24) shows the second assumption is probably violated.

In each of the proposed models, collinearity is present in such magnitude that it's presence can't be ignored. pages 21, 23 and 25 give the model variables, p-values and VIF's associated with each of the previously mentioned models.

NEW MODELS FOR LOADTIME DATA

I have attempted to improve on past modeling endeavors made by PRC witnesses. An improved model will retain a high R_a^2 statistic, provide reasonable coefficients, not grossly violate model assumptions, and possess other qualities compatible with a good linear model to be used for estimation purposes. I will present a general overveiw of my model building procedure that was used for all three data sets. Details of my model building procedure will be presented using the BAM data set as an example.

Each of the three data sets contains a large number of potential regressor variables. There are p^* regression variables generated for each delivery site which form the pool of regressor variables. For the BAM, MDR and SDR delivery sites, $p^* = 226$, 239, and 182 respectively.

Due to a small number of observations (usually zero) and other considerations, the set of regressor variables, although similiar, is unique for each delivery set. Since I'm working with such a large parameter set, it will be necessary to trim some of the less important variables from each data set.

A sequential model building procedure will be used to reduce the number of potential regressor variables from each data set. There are three basic sequential model building procedures: 1) forward selection, 2) backward selection, and 3) stepwise selection. With the forward selection model building procedure, the model starts out empty and variables are added to the model based on the significance of their respective partial F statistics. Once a variable enters the model, it remains in the model through out the model building procedure. With the backward model building procedure, all regression variables are fit to the model, and then variables are deleted based on the insignificance of their respective partial F statistics. Once a variable leaves the model, it never re-enters the model. With the stepwise model building procedure, variables may enter, exit, re-enter, and re-exit the model in a series of steps. Again, the significance of partial F statistics will determine which variables stay or are deleted from the model. I have used the stepwise model building procedure and will elaborate on it.

The stepwise procedure (PROC STEPWISE) available in Statistical Analysis Systems© (SAS) will be employed, as an exploratory tool, to obtain a preliminary subset of regressor variables. PROC STEPWISE is a sequential model building procedure. Typically, the model starts out empty, and then variables are added to or deleted from the model through a series of steps. In the first step of PROC STEPWISE, a regression model is fit for each of the p^* regressor variables. A set of partial F statistics (F_i^*, i=1,..,p^*) is calculated. The set contains F_i^* for each X_i.

$$F_i^* = \frac{MSR(X_i)}{MSE(X_i)}, \ i = 1, ..., p^*$$

The variable with the largest F_i^* is a candidate for addition to the model. This variable must be significant at some predetermined alpha level. In SAS this is called the significance level entry (SLE). The default SLE value is 0.15, but it can be changed (I did change the default values and will explain more about this later). A variable will not enter the model unless it meets the SLE criterion.

At the second step, a new set of partial F statistics (F_j^{**} j≠i) is calculated. This set is different than the original set of partial F statistics.

$$F_j^{**} = \frac{MSR(X_j|X_i)}{MSE(X_i, X_j)}, \ forj \neq i$$

The numerator of F_j^{**} is the mean square due to regressing X_j given the model already contains X_i. The demoninator of F_j^{**} is the mean square error of the model containing X_i and X_j. Again, the variable with the largest F_j^{**} is a candidate for addition to the model, and it must meet SLE criterion. After the second variable is added, the first added variable is rechecked for significance based on a model containing two terms.. A similiar F_j^{**} is calulated, where X_i takes the place of X_j. The first added variable remains in the model if a second significance level is met, and in SAS this is called the signifcance level stay (SLS). The default SLS value is 0.15.

In each successive step, the variable with the largest partial F statistic, given the model already contains those terms from the previous step,. is added to the model. Each added variable must meet SLE criterion, and all variables from the previous step must meet SLS criterion. It is important to note that variables continually enter and exit the model based on the significance of their respective partial F statistics. These partial F statistics are derived as a ratio of two mean squared values. The observed value of the partial F statistic of a particular variable changes between steps. Hence, the significance of a particular variable will, almost surely, change between steps. The exchange procedure will terminate when no other variables can enter or exit the model based on the SLE and SLS criterion.

After PROC STEPWISE has run, a rather large file (.lis output file) exists containing information pertinent to each step in the procedure. A list of associated model variables, as well as some summary statistics, are generated for each step. At the end of the .lis output file is a short summary for each step in the procedure. Here we find the variable that has entered the model, and the variable that has exited the model with respect to each step.

Two other important pieces of information are given for each step: the degress of freedom (d.f.), and Mallow's C_p statistic (C_p) where

$$C_p = p + \frac{(s^2 - \hat{\sigma}^2)(n-p)}{\hat{\sigma}^2}$$

Here, p=the number of model parameters with respect to the particular step, n=number of observations, s^2 = mean square error based on the p terms in the model, and $\hat{\sigma}^2$ is equal to the overall mean square error which is calculated based on the complete set of terms available to PROC STEPWISE. $C_p = p$ implies one of two things: 1) $p = n$ or 2) s^2=$\hat{\sigma}^2$. The first implies a parameter estimate has been found for each of the n observations (a silly thing to have). The second implies all the error in $\hat{Y}$ is variance (a good thing to have). C_p can be used as a criterion for selecting one model over another. A model that does not contain estimated bias will be one that has $C_p = p$ (Myers 1986). Since p is an integer and C_p probably is not we must look at those models where C_P is approximately equal to p. Since $p =$ d.f.+1, the variables associated with the step where C_p approximately equal to d.f. + 1 will be used as the *suggested* subset with respect to the stepwise procedure. The suggested subset of regressor variables will be used in the regression procedure available in SAS (PROC REG), and the model derived from PROC REG will be subjected to a diagnostic technique which further reduces the number of regressor variables. The goal is to find a subset of regressor variables that are significant, and multicollinearity among the regressor variables is not an obvious problem. The p-values of each regressor variable are included as part of the usual .lis output file, and will be used to establish the significance of the variables. The problem of multicollinearity will be handled through an inspection of VIF's. The VIF's can be made available to the analyst through PROC REG.

The VIF's are included as part of the .lis output file by inserting "VIF" at the end of the model statement in PROC REG. All variables that are associated with large VIF's should be deleted from the model or perhaps an alternative to least squares should be considered (Myers 1986). I will consider variable deletion as a possible solution to the collinearity problem. The term with the largest VIF should be deleted first, and then PROC REG is re-run . The .lis output file is re-examined, the variable with the largest VIF is removed, and PROC REG is re-run. This produces a new type of "backward" model building procedure (new in that it is very much different than the traditional backward model selection procedure). However, the deletion of one variable will, in general, affect the VIF of another variable or variables. Careful consideration is needed in determining which variables should be deleted based on inspection of VIFs. When comparing VIF's, variables with the largest VIF's should be compared to other terms in the model. Deleting a suspect term in the model will probably lower the VIF's for some of the other variables, and it can affect their respective significance levels as well. I suggest a careful analysis which involves looking for related terms with high VIF's, and deleting the highest order term first. Then, through a combination of significance at the 0.05 level and inspection of VIF's, other variables are deleted. The final result is a model where all terms are significant at the 0.05 level, and collinearity between the regressor variables is not an obvious problem. Hence, this should be a good model for estimation purposes. Further diagnostics of the model will be provided with a Q-Q plot (using studentized residuals), and a plot of t_i vs $\hat{y}_i$.

If the Q-Q plot is approximately linear with a slope of of one, and no obvious pattern is present in the plot of t_i vs $\hat{y}_i$ then it is reasonable to assume the second assumption has not been violated. However, linear regression is sensitive, in a detrimental way, to the presence of outliers. If this assumption appears to be violated, due to the presence of a few extreme observations, a reweighted least squares approach might be attempted. The reweighted least squares may produce a more robust model. It has been suggested that some data points appear to have been recorded in error due to unusally high recording times so that variable deletion might be a reasonable solution. If a reweighted least squares approach is used, another Q-Q plot will be generated to check the normality assumption. To this end, a parsimonious model may be found such that: 1) the estimated regression coefficients are sensible and significant, 2) collinearity among the regressor variables is not an evident problem, and 3) model assumptions are not grossly violated. Hopefully, this model also exhibits a relatively high R_a^2.

AN EXAMPLE: MODELLING THE BAM DATA

The BAM data set consists of 1442 observations recorded on each of sixteen variables. For the regression analysis, three variables form the response variable DTSUM (defined earlier) and seven form the original explanatory variables (two of which are qualitative). The quantitative explanatory variables are: accountables delivered (AD), letters delivered (LD), flats delivered (FD), parcels delivered (PD), and volume collected (VC). Product terms for LD, FD, and VC have been constructed, as well as, dummy variables for each of the terms that make up DTSUM. The two qualitative variables are for receptacle and container code types. Receptacle type has eleven categories, and container code has six categories.

It has been suggested that some components of load time may have been recorded in error, as signified by unusually high recording times; this point was referred to the PRC. Although the recorded values for some of the data points appear questionable, I, personally, could not disqualify the assumption that the data were recorded without error. A regression model will be fit to those observations available for analysis.

To use the qualitative variables in a regression analysis, a quantitative dummy variable is generated for each of the receptacle and container code types. Product and interaction terms are formed between the "new" and the original quantitative variables. A series of data manipulation steps in SAS provides a means for building the pool of regression variable. The complete pool of regression variables actually consists of 226 variables. For the most part, the data will remain as recorded. However, due to a small number of observations in receptacles 6 and 7 (MR6 and MR7), their observations have been dumped into receptactle 11 (MR11). MR11 is an "other" category, presumably used when a receptacle does not fall into one of the more well-defined categories. Likewise, observations in container code 5 (CT5) are dumped into container code 6 (CT6). The combining of some observations was suggested by more experienced personnel. The complete data set is now available for use in PROC STEPWISE.

Again, PROC STEPWISE will be used to obtain a smaller subset of regressor variables from the pool of possible regressor variables. It has been suggested (by more experienced personnel) that 27 variables should be forced into the model for the entire stepwise procedure. Because, they are lower order terms, that are believed to be important, and hence should remain in the model as long as possible. That is, they will remain in the model until the final stages of model refinement, and then will be removed only if it is beneficial for the model. The 27 variables are forced into the model by using the "include option" available in SAS. Specifically, inserting "include 27" at the end of the model statement in SAS, will force the first 27 variables in the model statement. The forced in variables are: AD, PD, FD, VC, PDS, $(AD)^2$,$(PD)^2$, $(FD)^2$, $(VD)^2$, $(PDS)^2$, PDUMMY, ADUMMY, LDUMMY, MR1-MR5, MR8-MR11, CT1-CT4, and CT6. As mentioned before, the default SLE and

SLS values are 0.15, I have chosen SLE and SLS values of 0.20 for both significance levels. The reason for choosing these levels is that collinearity may be present among the regressor variables, and I want a good subset of the original set of regressor variables that possibly contains, as a subset of it, a set of significant variables (significant at the 0.05 level), where collinearity is not an evident problem.

The default SLE and SLS values are changed by inserting "SLE=.2 SLS=.2"at the end of the model statement in SAS. Since variables are forced into the model, and the SLE and SLS values have been changed, "include 27 SLE=.2 SLS=.2" is inserted at the end of the model statement in SAS to accomplish both of the above mentioned objectives. As PROC STEPWISE is run, variables will enter and exit the model until, at the final step, the model includes 27 forced in variables and those "unforced" variables that have met the SLE=.2 and SLS=.2 criterion.

After PROC STEPWISE has completed, an examination of the end of the .lis output file reveals that step where where C_p is approximately equal to the number of model parameters. Step 38 shows $C_p = 57.76748$, and the number of terms in the model equals 58 for this step. After locating step 38 in the main body of the .lis output file, the variables associated with step 38 become the subset of regressor variables *suggested* by PROC STEPWISE. The subset of regressor variables will be used in PROC REG for further model refinement. As a note, the R^2 statisic is reported as .8878, and R_a^2 is not provided as part of the usual .lis output file in PROC STEPWISE.

PROC REG is run with the VIF option using the 58 variables suggested by PROC STEP-

WISE. As a note, the initial run (run 1) of PROC REG yields: R^2= .8878, and R_a^2=.8830. R^2 is consistent with that obtained from PROC STEPWISE, and provides a check that the desired terms have been transferred to PROC REG. An examination of the VIF's reveals 10 unforced terms and 16 forced terms with VIF's over 10. The largest VIF is 40275, and it is associated with the variable LD2MR4. Variables CT4MR4 and PDS2 have VIF's over 10000. The variable LD2MR4 will be deleted, and PROC REG will be re-run. Run 2 yields: R_a^2 = .8810, and the largest VIF drops to 962 (from variable CT4MR4). The variable CT4MR4 will be deleted, and PROC REG will be re-run. Run 3 yields: R_a^2 = .8796, and the largest VIF now drops to 95 (from variable CCDL2). The variable CCDL2 will be deleted and PROC REG will be rerun. This process of deleting a variable and re-running PROC REG continues until all unforced variables have VIF's less than ten. From inspection of the .lis output file corresponding to run 9, it is found that the VIF's for all unforced variables are under 10.

Now, the focus of the model building procedure will shift, and insignificant, unforced variables will be deleted from the model. The variable CT6MR11 has a p-value of .6501 and a VIF of 1.65. This is the largest p-value associated with all unforced variables, and CT6MR11 will be the first variable to be deleted based solely on significance. If two variables have p-values that are approximately the same then the variable with the larger VIF will be deleted first. Run 12 reveals that all unforced variables have VIF's less than 10, and the largest p-value for an unforced variable is 0.06 (associated with LDMR2). As a note, there are 6 forced in variables that have VIF's over 10. The emphasis for variable deletion will shift again. Now, those variables (forced or unforced) that are least significant will be deleted one at a time. Variable MR10 has a p-value of .9322, and is the first candidate for deletion. Again, this process of deleting variables, one at a time, will continue until all variables are significant at the 0.05 level.

From an inspection of the .lis output file associated with run 27, it is found that all variables are significant at the 0.05 level. There is still a slight problem with VIF's. A few of the forced in variables still have VIF's over ten. Variables with the largest VIF's will be deleted until all variables are significant at the 0.05 level, and the VIF of no variable is over 10. Inspection of the .lis output file, associated with run 32, reveals that all variables are significant, and all VIF's are less than 10. The largest VIF is 7.7 (from ADUMMY), and the largest p-value is 0.0492 (from FD2MR1). The final subset of model variables has now been determined, and R_a^2 is found to be 0.847.

A Q-Q plot of t_i vs the ith order statistic of t_i has been generated using SAS. The plot does not follow the ideal (see page 35). There is an s-shape which suggests the error terms are heavily tailed. That is, a few observations have unusually large residuals. A plot of t_i vs. $\hat{y}_i$ is not very helpful (see page 34); a few extreme observations have compromised the effectiveness of the plot.

It has been suggested to perform a re-weighted least squares. From the last run of PROC REG, $\sqrt{MSE}$ is found to be 185.48, and 3*$\sqrt{MSE}$= 556. If ((-556 $\leq e_i \leq$ 556) , i=1,...,1412) is a false statement for observation i then y_i is given weight zero. Subsequently, the ith observation is deleted from the data set and the model is refit. After this refitting is complete, R_a^2 is equal to 0.943, and an inspection the Q-Q plot presented on page 36 shows it is reasonable to conclude the normality assumption has not been violated. Hence, the re-weighted least squares approach appears to have helped the model.

The estimated coefficients, associated VIF's, and p-values for the terms associated with each of the three final models are presented on pages 31, 32 and 33. As a note, the response variable is DTSUM for the MDR and BAM data sets, while a log transformation of the response (Ln(DTSUM)) was used for the SDR data set. The log transformation was suggested by more experienced personnel. On pages 27-30, tables summarizing the variable deletion process for each data set can be found. These tables show which variables were deleted, R_a^2, VIF's , p-values, and the criterion for deletion corresponding to each step. As an important note, preliminary data analysis found the SDR data set to be recorded in error. The data from observation sites 333 and 377 were entered in duplicate! This error was not discovered by the PRC witnesses. The model I fit to the SDR data excluded the duplicate observations.

Conclusion

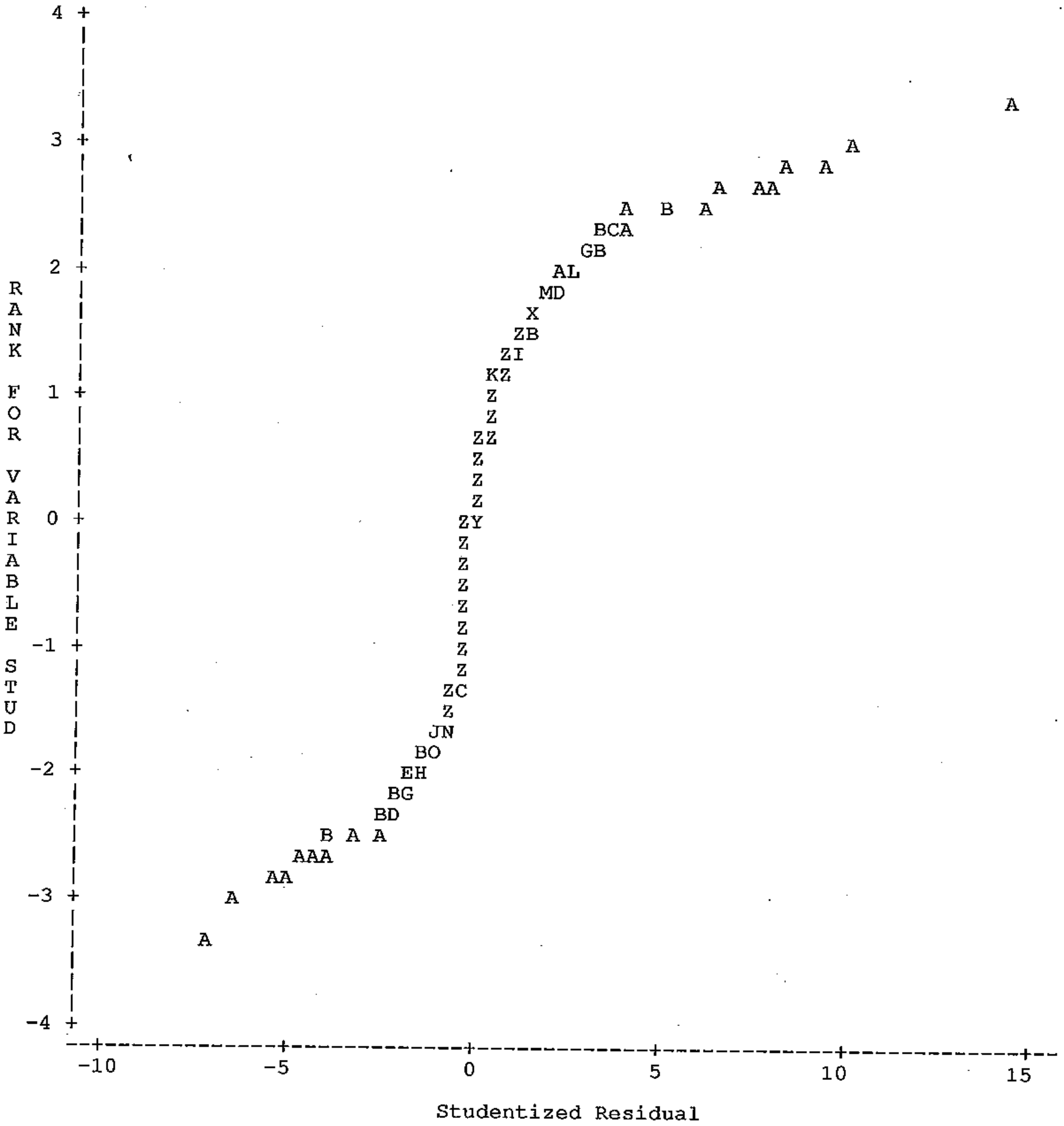
I have improved on the previously proposed loadtime variablilty models. The models are improved in that: 1) R_a^2 has increased for each model, 2) VIF's are reasonable in magnitude, thereby removing the collinearity problems, 3) the model variables are statistically significant, and 4) basic assumptions have been verified. The PRC has taken notice this model building procedure and will be considered as a viable procedure in future models of loadtime.

Diagnostics of PRC Proposed Models
BAM DATA

Variable	VIF	p-value
MR6	1.1	.0336
MR8	1.0	.0001
CT1	1.1	.0759
CT3	1.0	.0001
PD	5.2	.0001
AD	5.4	.0001
PDS	53.5	.0001
LD2	2.3	.0001
PD2	5.2	.0062
AD2	4.6	.0001
VC2	17.6	.0119
PDS2	14351.2	.0001
LVDC	36.6	.0217
LDPDS	11094.0	.0001
FADD	2.1	.0001
FVDC	10.4	.0839
FDPDS	296.2	.0001
ADPDS	4.5	.0005
VCPDS	77.9	.0001

Q-Q Plot for the PRC Model
BAMLTV DATA

Plot of ZSCORE*STUD. Legend: A = 1 obs, B = 2 obs, etc.



Diagnostics of PRC Proposed Models
MDR DATA

Variable	VIF	p-value
LD	13.5	.0444
FD	11.0	.0022
PD	3.6	.0001
VC	2.0	.0001
AD	2.5	.0001
PDS	11.6	.0001
FD2	67.4	.0001
PDS2	17.4	.0001
LFDD	92.4	.0001
LPDD	58.0	.0001
LADD	97.3	.0953
LVDC	22.5	.0001
FPDD	19.5	.0001
FADD	25.0	.0001
FVDC	7.6	.0001
FDPDS	53.7	.0402
PADD	44.2	.0002
PVDC	24.2	.0001
ADPDS	57.8	.0004
VCPDS	8.4	.0844
MR2	1.2	.0301
MR7	1.3	.0001
MR8	1.0	.0464

```
                            Q-Q Plot for the PRC Model
                                     MDR DATA

              Plot of ZSCORE*STUD.  Legend: A = 1 obs, B = 2 obs, etc.

     |
     |
   4 +
     |
     |
     |                                                                A
     |
   3 +                                                     A          A
     |                                                    AA
     |                                              AA B
     |                                            BCA
R    |                                          AGC
A  2 +                                         KD
N    |                                       HO
K    |                                      JV
     |                                      Z
F    |                                     ZI
O  1 +                                     Z
R    |                                    ZZ
     |                                    Z
V    |                                    Z
A    |                                    Z
R  0 +                                    Z
I    |                                    Z
A    |                                    Z
B    |                                   ZZ
L    |                                   Z
E -1 +                                   Z
     |                                   Z
S    |                                  ZK
T    |                                 MS
U    |                                HO
D -2 +                              AFH
     |                             CH
     |                            CC
     |                        A AB
     |                   B
  -3 +             A    A
     |
     |          A
     |
     |
  -4 +
     |
     -+-----------+-----------+-----------+-----------+-----------+-----------+-
     -15         -10          -5          0           5          10          15

                                 Studentized Residual
```

Diagnostics of the PRC Proposed Models
SDR DATA

Variable	VIF	p-value
MR2	1.4	.0001
MR3	1.5	.0001
MR4	1.0	.0001
MR5	1.1	.0008
MR7	1.0	.0001
MR8	1.1	.0001
MR9	1.1	.0139
CT1	3.0	.0001
CT3	2.0	.0001
CT4	2.2	.0001
LD	2.0	.0001
FD	1.7	.0001
PD	4.4	.0001
AD	2.4	.0001
VC	5.7	.0001
LD2	22.1	.0001
PD2	4.5	.0001
VC2	10.4	.0001
LFDD	9.2	.0001
LADD	4.7	.0001
LVDC	70.4	.0001
FADD	3.8	.0323
FVDC	19.5	.0623
PADD	2.3	.0001
PVDC	154.3	.0001
AVDC	159.2	.0001

```
                        Q-Q Plot for PRC Model
                               SDR DATA

        Plot of ZSCORE*STUD.  Legend: A = 1 obs, B = 2 obs, etc.

     |
     |
   4 +                                                               A
     |                                                 A
     |                                                AA
     |                                            A C
     |                                        BCAA
   3 +                                     BEEC
     |                                  EDIG
     |                                SOJ
     |                              PZI
R    |                             ZZ
A  2 +                            ZZ
N    |                            Z
K    |                           ZZ
     |                           Z
F    |                           Z
O  1 +                           Z
R    |                          ZB
     |                          Z
V    |                          Z
A    |                          Z
R  0 +                          Z
I    |                          Z
A    |                          Z
B    |                         ZZ
L    |                         Z
E -1 +                         Z
     |                         Z
S    |                         Z
T    |                         Z
U    |                        LZ
D -2 +                        Z
     |                        Z
     |                        Z
     |                       ZQ
     |                     AMK
  -3 +                    IF
     |                   FA
     |                 AC
     |              AA
     |              A
  -4 +      A
     |
     -+------------+------------+------------+------------+------------+
     -20          -10           0           10           20           30

                            Studentized Residual
```

Summary of the Variable Deletion Process
BAM DATA

Run Number	Variable Deleted	R_a^2	VIF	P-value	Deletion Criterion
1	LD2MR4	.8830	40297	.0001	VIF
2	CT4MR4	.8810	962	.0001	VIF
3	CCDL2	.8796	95	.0072	VIF
4	MRDL11	.8791	90	.0019	VIF
5	DUMPL	.8783	64	.0236	VIF
6	PDS2MR11	.8779	44	.0001	VIF
7	PDS2CT2	.8501	14	.4430	VIF
8	LDCT4	.8501	11	.0132	VIF
9	CT6MR11	.8496	1.65	.6501	P-VALUE
10	PDAD	.8497	2.44	.2313	P-VALUE
11	PDSMR8	.8496	1.59	.2320	P-VALUE
12	MR10	.8496	1.1	.9322	P-VALUE
13	MR4	.8497	1.9	.9126	P-VALUE
14	PD2	.8498	5.1	.7200	P-VALUE
15	MR3	.8499	1.3	.6970	P-VALUE
16	CT2	.8500	17.1	.6695	P-VALUE
17	VC2	.8501	4.8	.241	P-VALUE
18	VC	.8500	1.2	.9569	P-VALUE
19	CT3	.8501	1.4	.7306	P-VALUE
20	CT4	.8502	1.5	.3891	P-VALUE
21	MR5	.8502	1.9	.3130	P-VALUE
22	MR8	.8502	1.0	.2200	P-VALUE
23	MR1	.8502	1.4	.1683	P-VALUE
24	CT2MR1	.8501	1.2	.1115	P-VALUE
25	MR9	.8499	1.4	.1741	P-VALUE
26	MR2	.8498	1.7	.1142	P-VALUE
27	LDMR2	.8497	1.2	.2426	P-VALUE
28	PDS2	.8496	24.2	.0001	VIF
29	AD2	.8481	6.4	.0271	VIF
30	FD2	.8477	6.5	.0396	VIF

Summary of the Variable Deletion Process
MDR DATA

Run Number	Variable Deleted	R_a^2	VIF	P-value	Deletion Criterion
1	VCLD	.9616	10936	.0003	VIF
2	LFDD	.9612	256	.0001	VIF
3	LPDD	.9531	195	.0663	VIF
4	FD2MR4	.9527	117	.0010	VIF
5	PADD	.9525	39.72	.0001	VIF
6	PDSPD	.9515	51.7	.0001	VIF
7	PVDC	.9492	32.4	.0328	VIF
8	LD2MR2	.9490	322	.0003	VIF
9	LD2MR11	.9488	32	.0001	VIF
10	PDS2MR11	.9375	12.8	.8605	P-VALUE
11	VC2MR11	.9376	21.5	.5000	P-VALUE
12	FPDD	.9376	18.8	.1620	P-VALUE
13	ADCT1	.9376	25.4	.0136	VIF
14	FDMR4	.9373	12.7	.9618	P-VALUE
15	FD2MR7	.9374	7.5	.8226	P-VALUE
16	LVDC	.9374	18	.0031	VIF
17	PDMR4	.9371	11.6	.1100	VIF
18	LDPD	.9370	12.5	.0001	VIF
19	PD2	.9358	17	.3282	VIF
20	PDS2	.9358	11.3	.1434	VIF
21	AD2	.9358	14	.0194	VIF
22	LD2	.9356	13.6	.0001	VIF
23	FDPD	.9309	3.1	.7489	P-VALUE
24	VCMR11	.9310	2.2	.9978	P-VALUE
25	ADMR2	.9310	1.5	.9411	P-VALUE
26	MRDP4	.9310	1.7	.8400	P-VALUE
27	ADMR11	.9311	2.5	.7914	P-VALUE
28	FD2MR2	.9311	1.1	.2850	P-VALUE
29	LDUMMY	.9311	1.1	.9240	P-VALUE

Summary of the Variable Deletion Process
MDR DATA cont.

Run Number	Variable Deleted	R_a^2	VIF	P-value	Deletion Criterion
30	PDMR2	.9312	1.2	.1689	P-VALUE
31	MR7	.9311	4.1	.7317	P-VALUE
32	PDMR3	.9312	4.9	.1658	P-VALUE
33	VC2MR1	.9311	1.1	.0622	P-VALUE
34	CT1	.9310	1.8	.4114	P-VALUE
35	CT3	.9310	1.0	.357	P-VALUE
36	CT2	.9310	1.1	.2342	P-VALUE
37	AD	.9310	2.1	.2703	P-VALUE
38	FD	.9310	8.4	.2427	P-VALUE
39	FDMR7	.9309	1.6	.0708	P-VALUE

Summary of the Variable Deletion Process
SDR DATA

Run Number	Variable Deleted	R_a^2	VIF	P-value	Deletion Criterion
1	DUMPL	.4752	172	.0001	VIF
2	MRDL3	.4750	124.9	.0839	VIF
3	VCLD	.4749	108	.0001	VIF
4	CT3MR3	.4741	106.9	.0003	VIF
5	LDLD	.4737	70.4	.0519	VIF
6	VCCT1	.4736	48.9	.0501	VIF
7	PDLD	.4735	38.5	.0015	VIF
8	FD2CT3	.4732	8.7	.0200	VIF
9	CT1MR9	.4731	16	.0001	VIF
10	LDMR5	.4723	16	.0010	VIF
11	AD2	.4710	13.4	.0452	VIF
12	PDAD	.4709	2.2	.8997	P-VALUE
13	CT1	.4709	15.6	.0301	P-VALUE
14	CT2MR5	.4708	1.4	.7873	P-VALUE
15	VC2CT2	.4708	1.1	.3652	P-VALUE
16	LD2MR2	.4708	1.9	.2464	P-VALUE
17	CT3MR2	.4708	1.1	.2290	P-VALUE
18	FADD	.4708	2.6	.1747	P-VALUE
19	ADMR9	.4707	2.5	.1025	P-VALUE
20	FPDD	.4707	1.1	.1030	P-VALUE
21	CT3MR1	.4706	1.1	.0791	P-VALUE
22	CT3	.4706	4.0	.4552	P-VALUE
23	AD	.4706	11.1	.0001	VIF
24	MR9	.4696	2.9	.2573	P-VALUE
25	PDMR3	.4696	1.6	.0585	P-VALUE

Proposed Model for the BAM Data

Variable	VIF	p-value
LD	6.3	.0001
FD	2.9	.0010
PD	1.2	.0273
AD	6.3	.0001
PDS	2.8	.0001
LD2	6.5	.0001
CT1	1.7	.0001
PDUMMY	1.2	.0001
LDUMMY	1.7	.0001
ADUMMY	7.7	.0001
FDMR1	2.17	.0001
ADMR5	5.8	.0001
ADMR11	1.9	.0001
PDSMR11	1.4	.0001
FD2MR1	1.0	.0492
PDS2MR2	1.0	.0012
FDCT1	4.9	.0001
PDCT3	1.0	.0001
VCCT3	1.0	.0001
PDSCT2	1.2	.0001
FD2CT1	4.0	.0001
CCDA1	4.9	.0026
MRDP9	1.2	.0017
MRDA9	1.6	.0001
MRDA11	1.9	.0003
CT4MR11	1.6	.0001
DUMPA	1.9	.0073
FDAD	2.3	.0018

Proposed Model for the MDR Data Set

Variable	VIF	p-value
LD	7.5	.0001
PD	2.8	.0001
VC	6.7	.0482
PDS	5.6	.0001
FD2	1.5	.0001
VC2	8.7	.0001
MR1	1.3	.0003
MR3	2.3	.0001
MR4	2.2	.0001
MR8	2.5	.0162
MR9	1.1	.0033
ADUMMY	1.8	.0001
PDUMMY	1.2	.0394
PDSMR11	5.8	.0001
LDMR3	3.2	.0001
LDMR11	5.8	.0001
PDMR11	2.0	.0001
LD2MR7	2.2	.0001
VC2MR4	1.9	.0001
VCCT4	1.2	.0408
FD2CT4	1.2	.0001
MRDP7	1.4	.0001
CT2MR2	1.0	.0001
VCPD	2.8	.0001
LDAD	4.5	.0001
VCAD	3.6	.0001
FADD	2.1	.0008
AVDC	3.1	.0001

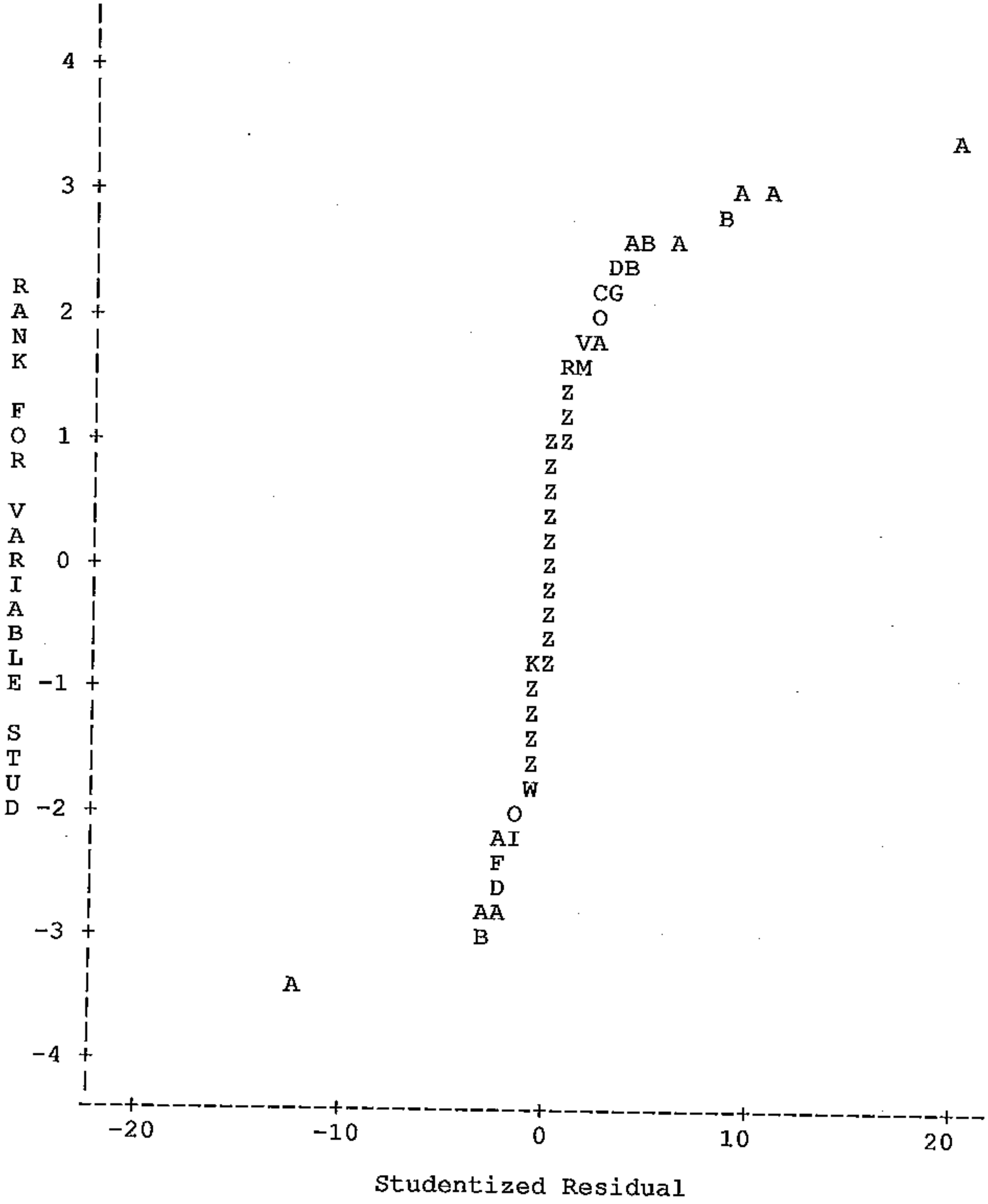
Proposed Model for the SDR Data

Variable	VIF	p-value	Variable	VIF	p-value
LD	2.8	.0001	LD2MR8	3.7	.0039
FD	5.1	.0001	FD2MR3	1.4	.0056
PD	5.0	.0001	FD2MR9	3.2	.0001
VC	4.1	.0001	VC2MR2	1.2	.0132
LD2	2.0	.0001	VC2MR3	1.2	.0003
FD2	2.9	.0001	LDCT3	2.0	.0217
PD2	4.4	.0003	FDCT3	2.0	.0015
CT2	3.3	.0001	PDCT4	1.1	.0013
CT4	1.1	.0001	VCCT4	1.0	.0020
MR1	1.6	.0001	MRDL5	4.8	.0001
MR2	3.9	.0001	MRDA8	1.1	.0001
MR3	5.4	.0001	MRDP2	1.8	.0001
MR5	5.4	.0001	MRDP3	3.1	.0001
MR8	1.6	.0001	MRDP5	1.3	.0060
PDUMMY	3.9	.0001	MRDP9	1.3	.0407
ADUMMY	3.2	.0001	CCDP2	1.5	.0001
LDUMMY	1.9	.0001	CCDA2	1.1	.0002
LDMR3	8.5	.0001	CT2MR2	1.1	.0001
FDMR2	2.8	.0001	DUMPA	2.0	.0207
FDMR9	3.3	.0001	FDPD	3.0	.0001
PDMR8	1.6	.00028	PDPD	2.0	.0323
PDMR9	1.2	.0001	ADPD	1.8	.0195
VCMR1	1.2	.0212	ADLD	1.9	.0023
VCMR8	1.2	.0106	ADAD	2.0	.0001
LD2MR3	1.1	.0001	FDAD	1.8	.0001

```
                    Plot of Student's t vs Predicted DTSUM

         Plot of STUD*PRED.  Legend: A = 1 obs, B = 2 obs, etc.

      |
      |
   20 +              A
      |
      |
      |
      |
      |
   15 +
      |
      |
      |
      |
S     |         A
t  10 +
u     |           A
d     |          A A
e     |
n     |
t     |               A
i   5 +           A
z     |         C A
e     |          C A
d     |        AJG
      |         SIHB
R     |         ZZGD  A
e   0 +         ZZOB A      A    A
s     |         ZZLDC
i     |          JKE A
d     |           GBA                                             A
u     |
a     |
l  -5 +
      |
      |
      |
      |
      |
  -10 +
      |
      |
      |                     A
      |
      |
  -15 +
      |
      -+--------+--------+--------+--------+--------+--------+--------+-
     -2500      0      2500     5000     7500     10000    12500    15000

                          Predicted Value of DTSUM
```

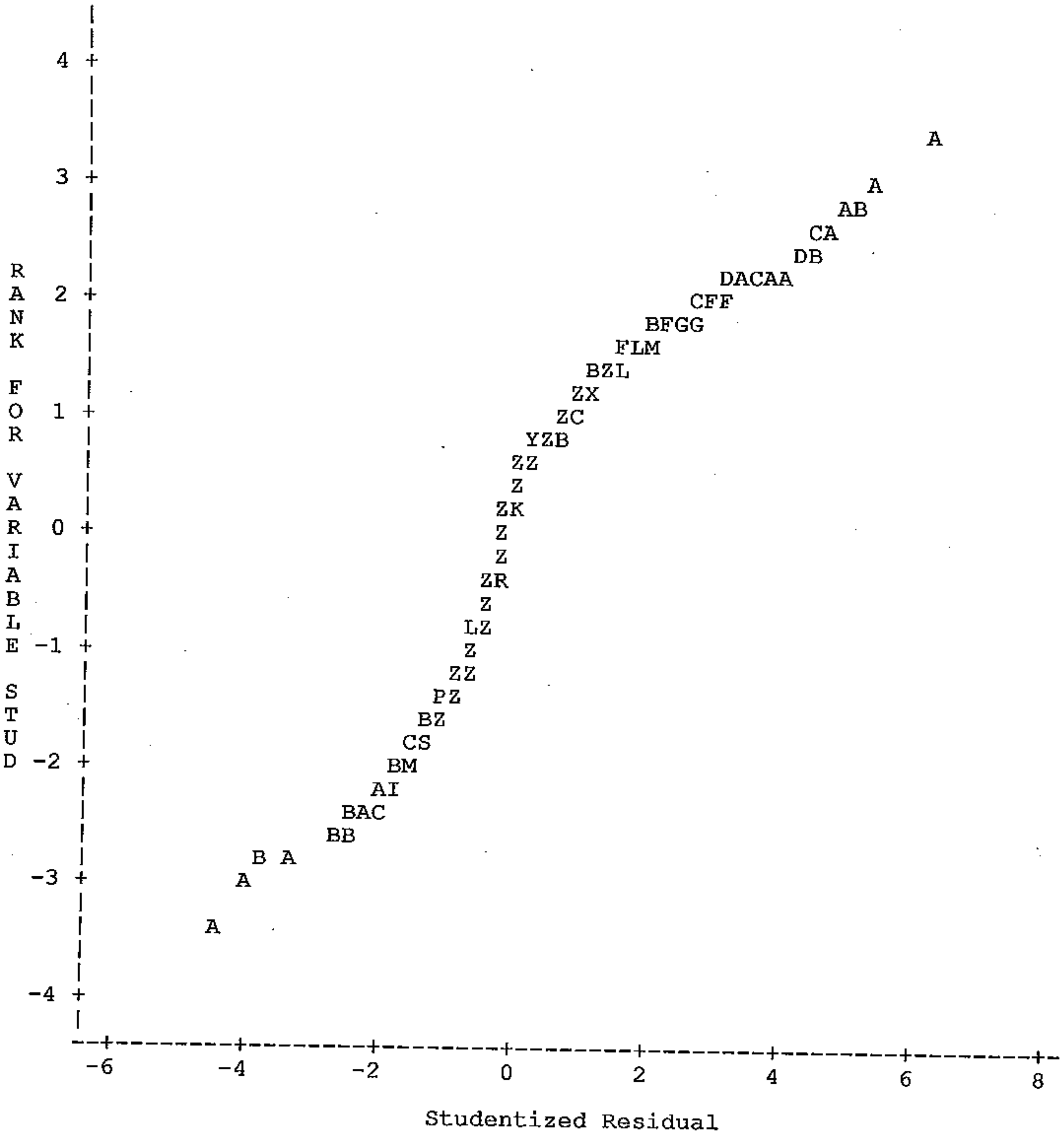
Q-Q Plot of Proposed Model
BAM DATA

Plot of ZSCORE*STUD. Legend: A = 1 obs, B = 2 obs, etc.



Q-Q Plot for the re-weighted BAM Model

Plot of ZSCORE*STUD. Legend: A = 1 obs, B = 2 obs, etc.



REFERENCES

1. Belsley, D.A., Kuh, E. & Welsch, R.E. (1980). *Regression Diagnostics.* New York: Wiley

2. Daniel, C. & Wood, F.S. (1980). *Fitting Equations to Data.* New York: Wiley.

3. Kennard, R.W. (1971). A Note on the C_p statistic. *Technometrics* 13: 899-900

4. Mallows, C.L. (1973). Some Comments on C_p. *Technometrics* 15: 661-675.

5. Myers, R.H. (1980). *Classical and Modern Regression with Applications* (2nd ed.). Boston: PWS-Kent Publishing Co.

6. Mosteller, F. & Tukey, J.W. (1977). *Data Analaysis and Regression.* Reading, MA: Addison-Wesley Publishing Co.

7. Neter, J., Wasserman, W. & Kutner, M. (1985). *Applied Linear Statistical Models* (2nd ed.). Homewood, IL: Richard D. Irwin

8. Postal Rate Commission Document (1989). *Cost Segment 7: City Delivery Carriers, Street Activity*: 7.0.2.

9. SAS Institute Inc. *SAS® User's Guide: Statistics*, Version 6.03. Gary, NC: SAS Institute Inc., 1985. 956 pp.

10. Seber, G.A.F. (1977). *Linear Regression Analysis.* New York: Wiley