Modeling of Survey Response Rates and Reporting Rates in the U.S. Consumer Expenditure Interview Survey

Moon J. Cho, John L. Eltinge
U.S. Bureau of Labor Statistics
Moon J. Cho, 2 Massachusetts Avenue NE, Washington, DC 20212
(Cho_M@bls.gov)

Key Words: Digit preference, Graphical methods, Interviewer effect, Model identification information and constraints, Model selection, Quarterly interview effect.

Abstract:

In the U.S. Consumer Expenditure (CE) Interview Survey, consumer units (roughly equivalent to households) are asked to provide month-by-month reports of the amounts of money spent on each of a large number of items. Reported expenditures are recorded at a relatively fine level of detail defined by the six-digit Universal Classification Code (UCC). For a given month, most consumer units report nonzero expenditures for a relatively small proportion of the possible UCC items. When no expenditure is reported, available data does not allow one to distinguish between cases of no expenditure for this item in the specified month ("true non-expenditure") and cases of failure to report a true non zero expenditure ("nonidentified item nonresponse"). However, under specific models for relationships among true non-expenditure, nonidentified item nonresponse and observable auxiliary variables, some important model parameters are estimable. This paper reviews the relevant models and available auxiliary information, discusses identifying restrictions for specific parameters, and presents related point estimators and variance estimators. The proposed methods are applied to selected subsets of items from the Consumer Expenditure Interview Survey.

1. Introduction

"Consumer expenditure (CE) surveys are specialized studies in which the primary emphasis is collecting data relating to family expenditures for goods and services used in day-to-day living." (BLS Handbook, 1997, p.160) One major use of the data is to provide the basis for revising weights and associated pricing samples for the Consumer Price Index (CPI). In addition, the BLS uses the data to produce estimates of mean expenditures and to produce public data sets of expenditures and income. The purpose of the CE Interview Survey is to obtain detailed data on relatively large expenditure items such as property, automobiles, or major appliances, or on expenses which occur on a fairly regular basis, such as rent, utility bills, and insurance premiums. The CE Interview Survey includes rotating panels: each consumer unit (CU) in the sample is interviewed every 3 months over five calendar quarters and then is dropped from the survey. Approximately 20 percent of the addresses are new to the Survey each quarter. The interviewer uses a structured questionnaire to collect both demographic and expenditure data in the Interview survey. See Cho et al.(2004) for more detailed information on the CE Interview Survey and related literature review.

2. Reporting Rates for CE Interview Survey Items

2.1 Aggregate and Interviewer-Level Estimators

Define J_{ic} as the number of non-zero reports obtained from a consumer unit c, by an interviewer i, and w_{ic} is the associated probability weight. Then a probability-weighted estimator of the overall proportion of nonzero reports is:

$$\hat{\pi} = \frac{\sum_{i=1}^{I}\sum_{c=1}^{n_i} w_{ic}J_{ic}}{\sum_{i=1}^{I}\sum_{c=1}^{n_i} w_{ic}J}$$

where I is the number of interviewers, n_i is the number of interviews conducted by an interviewer i, and J is the total number of item categories in our data. We also define $\hat{\pi}_{uw}$ as an unweighted estimator of

The views expressed in this paper are those of the authors and do not necessarily reflect the policies of the U.S. Bureau of Labor Statistics.

the overall proportion of nonzero reports:

$$\hat{\pi}_{uw} = \frac{\sum_i \sum_c J_{ic}}{\sum_i \sum_c J} \quad .$$

Define $\hat{\pi}_{i,uw}$ as an unweighted estimator of the proportion of nonzero reports for an interviewer i:

$$\hat{\pi}_{i,uw} = \frac{\sum_{c=1}^{n_i} J_{ic}}{n_i J} \quad .$$

Analysis on the coefficient of variation estimates of associated final weights for interviewers demonstrated that weights are generally homogeneous within interviewers. This implies that design effects are relatively small, and analysis on the unweighted reporting rate at the interviewer level is approximately equivalent to the weighted reporting rate.

2.2 Preliminary Study

For the preliminary study, we explored relationships between observed reporting rates and interviewer characteristics, interview number, reference month within reference period, and calendar month respectively (Cho et al., 2004). We noted that reporting rates for the fifth interview were slightly higher than the ones for three of the previous interviews. We observed that the average reporting rate for a supervisory field representative (SFR) tends to be substantially lower than those for a standard field representative. A possible explanation is that although SFR were considered to be the best interviewers, the cases they dealt with were harder ones. For standard field representatives, we explored the relationships between interviewer-level reporting rates and interviewer workload. We also observed that there were higher reporting rates in the more recent months. Consequently, the reporting rate was the highest for the month most recent to the time of the interview. We observed a significantly elevated reporting rate for December and a slightly elevated rate for August.

3. Estimation of Models for Reporting Rates

3.1 Logistic Regression Models

We consider the following logistic regression model based on initial exploratory analyses:

$$log(\frac{\hat{\pi}_{ir}}{1-\hat{\pi}_{ir}}) = \beta_0 + \beta_1 I_{i,r1} + \beta_2 I_{i,r2} \qquad (1)$$

where $I_{i,r1} = 1$ if the weighted reporting rate $\hat{\pi}_{ir}$ is from recall month 1 for an interviewer i; otherwise $I_{i,r1} = 0$. Similarly, $I_{i,r2}$ is an indicator for recall month 2, and n_i is the number of interviews conducted by interviewer i as previously defined. We fit the model for the combinations of interviewer i and recall month r in such a way that the number of observations to compute $\hat{\pi}_{ir}$ is greater than or equals to 5. We used weighted least squares to compute estimates $\hat{\beta}$ of the coefficients in the model, and estimated a covariance matrix for $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)$ using the balanced repeated replication method. The following table presents coefficient estimates, standard error estimates, and associated test statistics for Model (1):

Coeff	Est	S.E.	Test Statistic
B0	-1.839	0.006	-323.621
B1	0.088	0.004	22.531
B2	0.007	0.003	2.413

The cutoff point using the Bonferroni simultaneous inference method at $\alpha = 0.05$ is 2.499. Note that all coefficient estimators except the one for an indicator for recall month 2 are statistically significant according to this criterion.

We have fitted models with various predictor variables such as workload, calendar months, recall months, supervisory status. The findings are that: a coefficient estimator of the workload was not statistically significant in most models; a coefficient estimator of an indicator for August purchase was not statistically significant; a coefficient estimator of an indicator for recall month 2 was statistically significant in the models where an indicator for December was present; a coefficient estimator of interviewer status was statistically significant in the models where an indicator for workload was not present.

Figure 1 shows the plot of residual from Model (1) against a logarithm of interviewer's workload, Log(n_i). It displays a dotted line which represents weighted least square predictors, and a solid line for the locally weighted regression smoothing predictors (loess). In locally weighted regression smoothing, the nearest neighbors of each point are used for regression; the number of neighbors is specified as a percentage of the total number of points. This percentage is called the span; the span size used in Figure 1 is 2/3. Previously, a loess predictor showed that an interviewer's workload was somewhat nonlinear to residual. We transformed interviewer's workload by taking a logarithm to achieve a reasonable linearity. For some general background on these smoothing methods, see the section of MathSoft (1995, section 7.11). We also tried several power transformations of n_i; however, there were not much significant improvements compared to log transformation.

We chose Model (2) as our working model:

$$\begin{aligned} log(\frac{\hat{\pi}_{ir}}{1-\hat{\pi}_{ir}}) &= \beta_0 + \beta_1 ln(n_i) + \beta_2 I_{i,r1} + \beta_3 I_{i,r2} \\ &\quad + \beta_4 I_{ir,Aug} + \beta_5 I_{ir,Dec} + \beta_6 I_{ir,FR} \end{aligned} \quad (2)$$

The following table presents coefficient estimates, standard error estimates, and associated test statistics for Model (2):

Coeff	Est	S.E.	Test Statistic
B0	-1.981	0.057	-35.076
B1	0.010	0.013	0.780
B2	0.101	0.007	14.056
B3	0.024	0.007	3.289
B4	0.029	0.010	2.901
B5	0.156	0.017	9.296
B6	0.104	0.020	5.120

The cutoff point using the Bonferroni simultaneous inference method at $\alpha = 0.05$ is 2.836. Note that all coefficient estimators except the one for interviewer workload term are statistically significant according to this criterion.

4. Interviewer-Level Diagnostics

4.1 Preliminary Analysis of Design Effects

Evaluation of design effects can offer some additional insight into patterns of variability of estimated reporting rates. In general, if a point estimator $\hat{\theta}$ of θ is based on data from a complex sample, then its design effect is defined to be

$$deff(\hat{\theta}) = V_c(\hat{\theta}_c)/V_{SRS}(\hat{\theta}_{SRS})$$

where $V_c(\hat{\theta}_c)$ is the variance of $\hat{\theta}_{SRS}$, evaluated with respect to the complex sample design; $\hat{\theta}_{SRS}$ is a standard estimator of θ that one would use if one had collected data through a simple random sample; $V_{SRS}(\hat{\theta}_{SRS})$ is the design variance of θ_{SRS}, evaluated with respect to the hypothetical simple random sampling design. Thus, the overall reporting rate estimator $\hat{\pi}$ has a design effect equal to

$$deff(\hat{\pi}) = V_c(\hat{\pi}_c)/V_{SRS}(\hat{\pi}_{SRS})$$

where $V_{SRS}(\hat{\pi}_{SRS}) = \hat{\pi}_{uw}(1-\hat{\pi}_{uw})/\left(\sum_{i=1}^{I}\sum_{c=1}^{n_i}\sum_{r=1}^{3} J\right)$, $J = 190$ (the maximum number of possible items reported in our data) and $I = 590$ (the number of interviewers whose workload are greater than or equal to 5). For the CE Interview Survey example, an estimator of $deff(\hat{\pi})$ equals

$$V_{BRR}(\hat{\pi})/\left\{\hat{\pi}_{uw}(1-\hat{\pi}_{uw})/\left(\sum_{i=1}^{I}\sum_{c=1}^{n_i}\sum_{r=1}^{3} J\right)\right\}$$

where $V_{BRR}(\hat{\pi})$ is a weighted estimator of the design variance of $\hat{\pi}$ computed through balanced repeated replication with 44 half-sample replicates, and $\hat{\pi}_{uw} = \sum_{i=1}^{I}\sum_{c=1}^{n_i}\sum_{r=1}^{3} J_{icr}/\sum_{i=1}^{I}\sum_{c=1}^{n_i}\sum_{r=1}^{3} J$. For our data, $\widehat{deff}(\hat{\pi}) = 170.89$. This large design effects reflects very strong clustering effects within consumer units, as well as clustering effects within interviewers, relative to the variability one would expect to have from the mean of a simple random sample of $\left(\sum_{i=1}^{I}\sum_{c=1}^{n_i}\sum_{r=1}^{3} J\right)$ Bernoulli(π) random variables. Also, the consumer units covered by a given interviewer i generally all come from the same stratum and primary sample unit, and have approximately equal weights. Under those conditions, the design effect for $\hat{\pi}_{i,uw}$ arises from consumer unit-level clustering effects. An approximately unbiased estimator of this design effect is:

$$\widehat{deff}(\hat{\pi}_{i,uw}) = V_c(\hat{\pi}_{i,uw}) \Big/ \left\{\hat{\pi}_{i,uw}(1-\hat{\pi}_{i,uw})/\left(\sum_{c=1}^{n_i}\sum_{r=1}^{3} J\right)\right\}$$

where

$$\begin{aligned} &V_c(\hat{\pi}_{i,uw}) \\ &= (\sum_{r=1}^{3} n_i)^{-1}(\sum_{r=1}^{3} n_i - 1)^{-1} \\ &\quad \times \sum_{c=1}^{n_i}\sum_{r=1}^{3}(J^{-1}J_{icr} - \hat{\pi}_{i,uw})^2 \end{aligned}$$

is a standard cluster-based estimator of the variance of a sample mean (Cochran, 1977, p. 65), and $\hat{\pi}_{i,uw} = \sum_{c=1}^{n_i}\sum_{r=1}^{3} J_{icr}/\sum_{c=1}^{n_i}\sum_{r=1}^{3} J$. Figure 2 presents scatterplot of $\widehat{deff}(\hat{\pi}_{i,uw})$ against interviewer-level reporting rate, $\hat{\pi}_{i,uw}$. Note that Figure 2 suggests a moderate positive association between $\widehat{deff}(\hat{\pi}_{i,uw})$ and $\hat{\pi}_{i,uw}$. To study this further, we used ordinary least squares to estimate the coefficients of the model,

$$\widehat{deff}(\hat{\pi}_{i,uw}) = \gamma_0 + \gamma_0\hat{\pi}_{i,uw} + e_i \ .$$

The resulting coefficient estimates were $\hat{\gamma}_0 = 0.33$, and $\hat{\gamma}_1 = 39.96$. Due to the limitations on within interviewer sample sizes the direct cluster-based variance estimators $V_c(\hat{\pi}_{i,uw})$ were unstable for some interviewers i. Consequently, interviewer specific diagnostics will use the following design-effect-based estimators:

$$V^*(\hat{\pi}_{i,uw}) = (\gamma_0 + \gamma_1 \hat{\pi}_{i,uw}) \left\{ \hat{\pi}_{i,uw}(1 - \hat{\pi}_{i,uw}) / \sum_{c=1}^{n_i} \sum_{r=1}^{3} J \right\} .$$

4.2 Interviewer-Level Standardized Residuals

Define $\hat{r}_i$ to be a standardized simple residual for an interviewer i :

$$\hat{r}_i = \frac{\hat{\pi}_{i,uw} - \bar{\bar{\pi}}}{\{V^*(\hat{\pi}_{i,uw})\}^{1/2}}$$

where $\bar{\bar{\pi}}$ is the average of $\hat{\pi}_{i,uw}$, and $V^*(\hat{\pi}_{i,uw})$ is given in the previous section. Figure 3 displays a quantile-quantile plot of the ordered standardized remainder terms against the corresponding $(i - 0.5)/I_{(5)}$ quantiles of the standard normal distribution, where $I_{(5)}$ equals the number of interviewers with $n_i \geq 5$. Under regularity conditions, if all of the interviewer-level sample reporting rates had the same expectation, and if the variance function model provided an adequate approximation to the true variance of the deviations $\hat{\pi}_{i,uw} - \bar{\bar{\pi}}$, then the standardized remainder terms would follow approximately a standard normal distribution. Then following standard practice in the literature on linear models, the normal quantile-quantile plot in Figure 3 should have its points arranged along a solid line with a slope of 1, and an intercept of 0. Consequently, substantial deviation from this solid line may indicate that the assumption of equal expectations of the $\hat{\pi}_{i,uw}$ may be problematic. In addition, because each point corresponds to an individual interviewer, substantial deviation of a given point from the dotted-line pattern may indicate that the corresponding interviewer i has a sample reporting rate $\hat{\pi}_{i,uw}$ that is unusually high or low, even after accounting for sampling variability and implicit multiple testing. Quantile plots can offer insight into aggregate patterns of model fit and can also help one identify specific interviewers who may be appropriate candidates for follow-up study. First, note that a dotted-line fit to the plotted points in Figure 3 had an intercept approximately equal to zero, but had a slope substantially greater than one. Since the denominator term $V^*(\hat{\pi}_{i,uw})$ was intended only to estimate the sampling error variance of $\hat{\pi}_{i,uw}$, the large slope indicates that additional error terms (e.g., lack of fit in the constant-expectation model $E(\hat{\pi}_{i,uw}) = \pi$, say) may make substantial contributions to the differences $\hat{\pi}_{i,uw} - \bar{\bar{\pi}}$. Second, note that there is substantial downward curvature in the left-hand side of the plot, indicating substantial left-skewness in distribution of the differences $\hat{\pi}_{i,uw} - \bar{\bar{\pi}}$. In other words, relative to the pattern one would expect with a Gaussian distribution, we have an unusually large number of interviewers with sample reporting rates $\hat{\pi}_{i,uw}$ substantially below the overall mean $\bar{\bar{\pi}}$. It may be of special interest to consider in greater depth the interviewers with $\hat{\pi}_{i,uw}$ values in the extreme parts of this distribution. It would also be of interest when we study digit preference (Swanson et al., 2003).

Define:

$$\widetilde{deff}(\hat{\pi}_{i,uw}) = V_c(\hat{\pi}_{i,uw}) / \left\{ \tilde{\pi}_i(1 - \tilde{\pi}_i) / \sum_{c=1}^{n_i} \sum_{r=1}^{3} J \right\}$$

where $\tilde{\pi}_i$ is weighted average of $\tilde{\pi}_{ir}$, and $\tilde{\pi}_{ir}$ is the fitted value for $\hat{\pi}_{ir}$ under Model (2),

$$\tilde{\pi}_i = (n_{i1}\tilde{\pi}_{i1} + n_{i2}\tilde{\pi}_{i2} + n_{i3}\tilde{\pi}_{i3}) / (n_{i1} + n_{i2} + n_{i3})$$

where $\tilde{\pi}_{ir} = e^{\tilde{y}_{ir}}/(1 + e^{\tilde{y}_{ir}})$, $\tilde{y}_{ir} = \hat{\beta} x_{ir}$, $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4, \hat{\beta}_5, \hat{\beta}_6)$ and $x_{ir} = \{1, ln(n_i), I_{i,r1}, I_{i,r2}, I_{ir,Aug}, I_{ir,Dec}, I_{i,FR}\}$. Then, using ordinary least squares to estimate the coefficients of the model,

$$\widetilde{deff}(\hat{\pi}_{i,uw}) = \tilde{\gamma}_0 + \tilde{\gamma}_1 \tilde{\pi}_i + e_i \ .$$

The resulting coefficient estimates were, $\tilde{\gamma}_0 = -5.60$ and $\tilde{\gamma}_1 = 87.68$. Finally, define

$$\widetilde{V}(\hat{\pi}_{i,uw}) = (\tilde{\gamma}_0 + \tilde{\gamma}_1 \tilde{\pi}_{i,uw}) \left\{ \tilde{\pi}_{i,uw}(1 - \tilde{\pi}_{i,uw}) / \sum_{c=1}^{n_i} \sum_{r=1}^{3} J \right\} .$$

In addition, define the standardized logistic regression remainder term

$$\widetilde{r}_i = \frac{\hat{\pi}_{i,uw} - \tilde{\pi}_i}{\{\widetilde{V}(\hat{\pi}_{i,uw})\}^{1/2}} \ .$$

Figure 4 displays the associated standard normal quantile-quantile plots, where we have excluded interviewers i who had $n_i \leq 5$. The general interpretation of Figure 4 is similar to that for Figure 3: points that deviate substantially from the solid line with slope=1 and intercept=0 correspond to interviewers with values of $\hat{\pi}_{i,uw}$ that are unusually high or low, relative to Model (2).

5. Discussion

This paper has explored some relationships among nonresponse and predictors like interview number and interviewer workload in the CE Interview Survey. The preliminary results presented here suggest some association of interviewer-level item reporting rates, with these predictors, and with data-quality measures defined by leading-digit prevalence rates. It would be of interest to consider extensions of this work in several area. For example, one could explore the extent to which reporting rates are associated with other indicators of data quality, or with measures of interviewer training and experience. In addition, one could consider the use of more detailed modeling and estimation methods.

6. References

Cho, M., Eltinge, J., and Swanson, D. (2003). Inferential Methods to Identify Possible Interviewer Fraud Using Leading Digit Preference Patterns and Design Effect Matrices. *2003 Proceedings of the American Statistical Association.* Survey Research Methods Section [CD-ROM], Alexandria, VA: American Statistical Association.

Cho, M., Eltinge, J., and Steinberg, B. (2004). Temporal Patterns of Survey Response Rates and Reporting Rates in the U.S. Consumer Expenditure Interview and Other Panel Surveys. *2004 Proceedings of the American Association for Public Opinion Research.* Survey Research Methods Section [CD-ROM], Alexandria, VA: American Statistical Association. To appear.

Cochran, W. (1977). *Sampling Techniques.* Third Edition, New York: John Wiley & Sons, Inc.

Eltinge, J., and Cho, M. (2003). Nonidentified Item Nonresponse and Reporting Rate in the U.S. Consumer Expenditure Interview Survey. *Proceedings of the 14th International Workshop on Household Survey Nonresponse.* Belgium. Available at http://www.nonresponse.org

Gabler, S., Haeder, S., and Lahiri, P. (1999). A Model Based Justification of Kish's Formula for Design Effects for Weighting and Clustering, *Survey Methodology*, vol. 25, pp. 105-106, 1999.

Groves, R., and Couper, M. (1998). *Nonresponse in Household Interview Surveys.* New York: John Wiley & Sons, Inc.

MathSoft, Inc.(1995), S-PLUS Guide to Statistical and Mathematical Analysis, Seattle, WA

Silberstein, A., and Jacobs, C. (1989). Symptoms of Repeated Interview Effects in the Consumer Expenditure Interview Survey. In D. Kasprzyk, G. Duncan, G. Kalton and M.P. Singh (eds.),pp 289-303, New York: John Wiley & Sons, Inc.

Silberstein, A. (1990). First Wave Effects in the U.S. Consumer Expenditure Interview Survey. *Survey Methodology*, Vol.16, No. 2, pp. 293-304.

Silberstein, A., and Scott, S. (1991). Expenditure Diary Surveys and Their Associated Errors. In P. Biemer, R. Groves, L. Lyberg, N. Mathiowetz, and S. Sudman (eds.), *Measurement Errors in Surveys*, pp 303-326, New York: John Wiley & Sons, Inc.

Spencer B. (2000). An Approximate Design Effect for Unequal Weighting When Measurement May Correlate With Selection Probabilities. *Survey Methodology*, Vol.26, No. 2 (December, 2000), pp. 137-138.

Swanson, D., Cho, M., and Eltinge, J. (2003). Detecting possibly fraudulent or error-prone survey data using Benford's Law. *Proceedings of the American Statistical Association.* Survey Research Methods Section [CD-ROM]. Alexandria, VA: American Statistical Association.

Wolter, K. (1979). Composite Estimation in Finite Populations. *Journal of the American Statistical Association*, Vol. 74, No. 367. (Sep., 1979), pp. 604-613.

Wolter, K. (1985). *Introduction to Variance Estimation.* New York: Springer-Verlag.

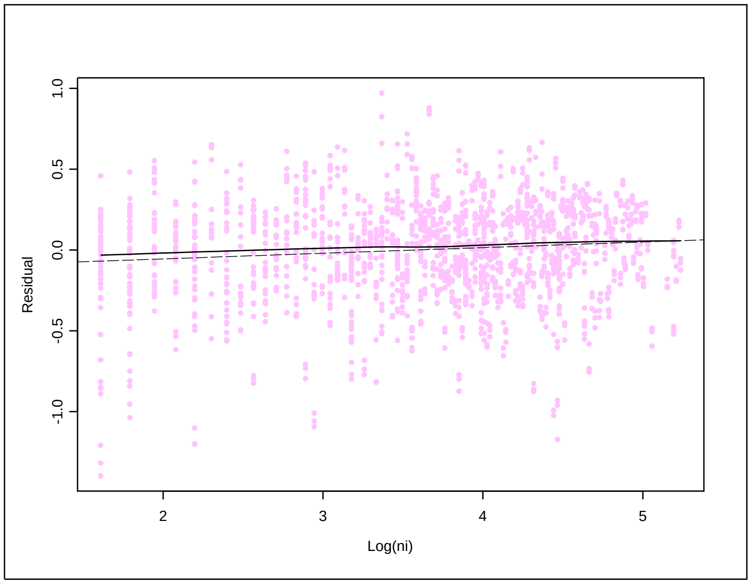

Figure 1: Residual against Log(ni)

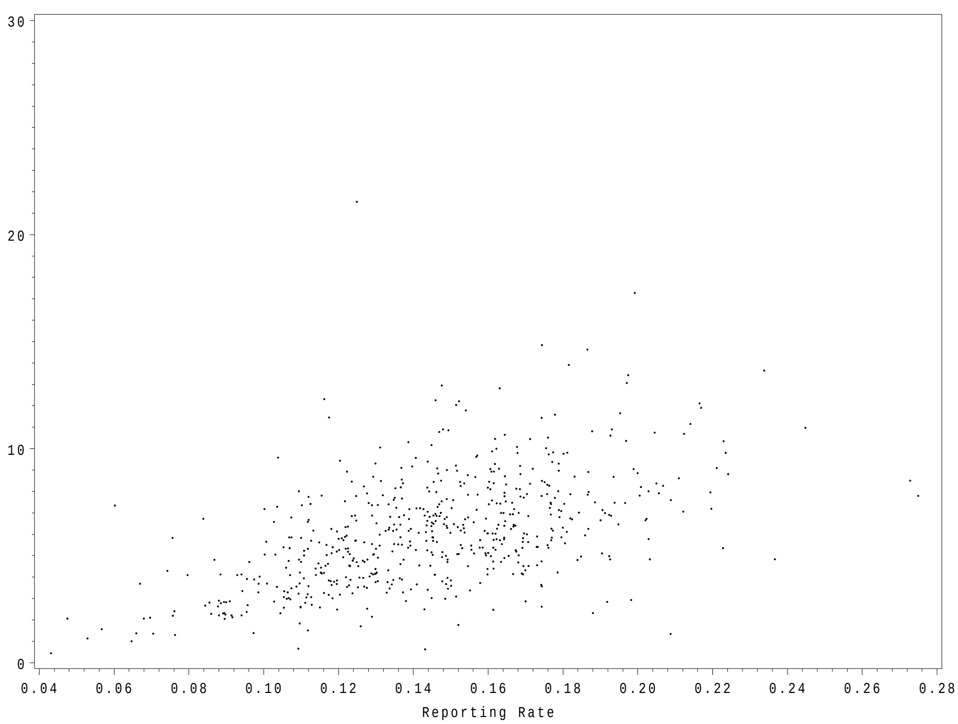

Figure 2: Plot of Deff against Reporting Rate ($\hat{\pi}_{i,uw}$)

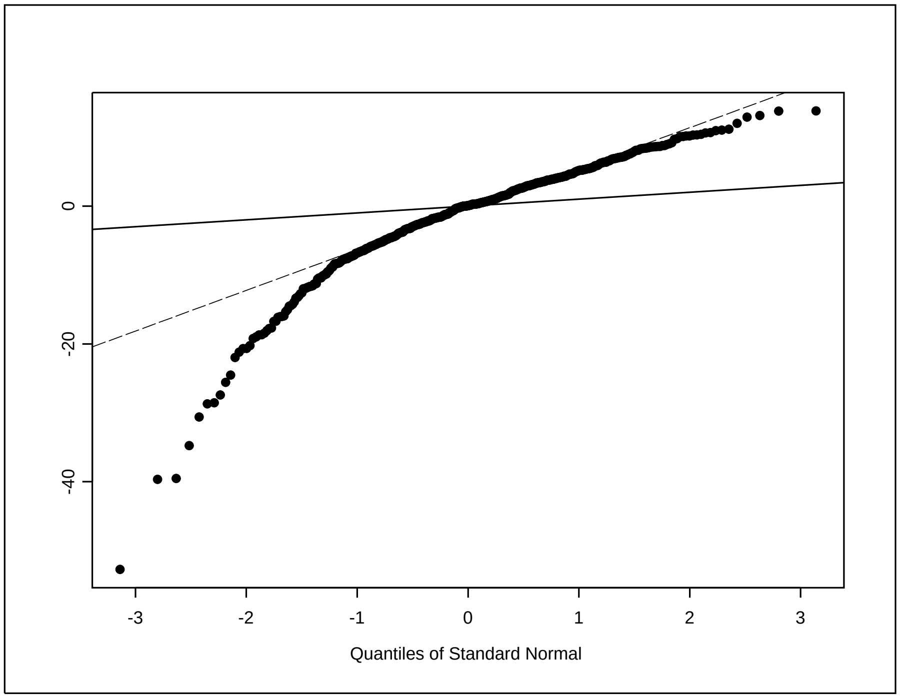

Figure 3: QQ Plot of Standardized Residual ($\hat{r}_i$)

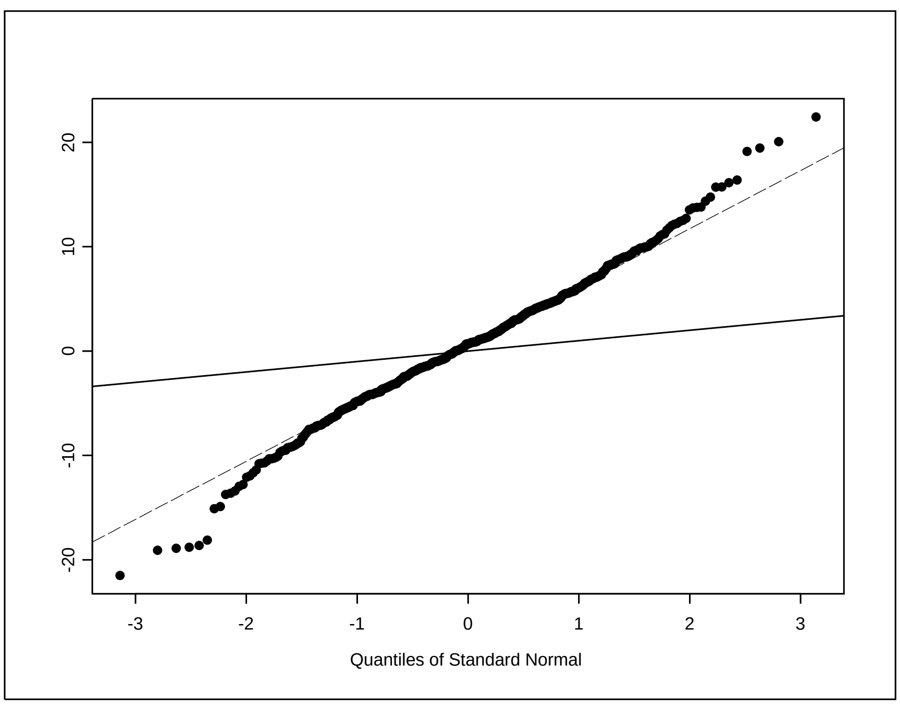

Figure 4: QQ Plot of Standardized Residual ($\widetilde{r}_i$)