Multivariate Regression Imputation Approach to the Analysis of Item Nonresponse in a Retail Trade Survey Data

Danhyang Lee[1]

Abstract
We present the result of regression imputation analysis in a retail trade survey data. The data have two industries with two study variables subject to missingness. The goal is to estimate the population totals for the two study variables using a suitable imputation method. We apply the deterministic multivariate regression imputation and then use the bootstrap procedure for variance estimation discussed by Shao and Sitter (1996). This takes into account for the imputation effect in addition to the sampling variance. In setting up the multivariate linear models, we divide data into sub-groups so that they have homogeneous variance structure within each group. The resulting confidence intervals cover the true population totals for each study variable.

Key Words: Multivariate linear model, Item nonresponse, Deterministic regression imputation, Bootstrap procedure

1. Introduction

The simulated data obtained from two industries(Industry XXX1 and XXX2) in a monthly retail trade survey are considered in this study. The data are incomplete due to item nonresponse. In each industry, a stratified simple random sample was obtained without replacement with six strata: one certainty(take-all) and five noncertainty strata. For each industry, the two study variables, current month sales(Sales00) and inventories(Inven00), are subject to missingness. The four covariates are current month administrative data value for sales(Asales00), prior month sales(Sales01), current month administrative data value for inventories(Ainven00), and prior month inventories(Inven01). Table 1 and Table 2 show their summary statistics for Industry XXX1 and XXX2, respectively. Note that Industry XXX2 has an unrealistic value of Ainven00 within stratum 2 which is negative. For our data analysis, this is replaced with the second smallest value of Ainven00 within stratum 2.

Based on the basic exploratory data analysis to be presented in Section 2, the imputation model is to be chosen to reflect the following observations: 1) Two study variables have distinct linear mean structure with other covariates; 2) Two study variables have moderate positive correlation; 3) Two study variables are subject to missingness with no partial missingness. In order to achieve our goal to estimate each total of the two study variables under this situation, multivariate regression imputation is put forward. Nonparametric imputation like nearest neighborhood approach might be considered. However, in this special case of our data having strongly linear relationship among variables, it is unlikely to beat a parametric imputation based on the linear regression models.

Before assuming multivariate regression models, we divide data into sub-groups so that it can have homogeneous variance structure (i.e., homoscedasticity) within each group. We do not assume (bivariate) normality because the given data is strongly right-skewed distributed.

[1]Danhyang Lee, Department of Statistics, Iowa State University, Ames, IA 50011, U.S.A., email: danhyang@iastate.edu.

We also provide variance estimates for each total after imputation by using the bootstrap procedure discussed by Shao and Sitter (1996). The Shao-Sitter (SS) bootstrap method provides asymptotically valid bootstrap estimator for imputed survey data, irrespective of the sampling design, imputation method or type of statistic used in inference. It can be obtained by imitating the process of imputing the original data set in the bootstrap resampling, and hence, takes into account the imputation effect.

This study is organized as follows. Section 2 describes data, how to apply the multivariate regression imputation and SS bootstrap method for variance estimation, which are two key methods applied to this study. Section 3 and Section 4 present some results and concluding remarks, respectively.

2. Application of Multivariate Regression Imputation to the Retail Trade Survey Data

We discuss data analysis for the two simulated data. Since using the same procedure for each simulated data due to their similar data patterns, we describe it only for Industry XXX1 in detail.

2.1 Data

Figure 1 and Figure 2 present two scatter plot matrices for the first five strata and sixth stratum of Industry XXX1 and XXX2, respectively. All of the scatter plots show positive linear patterns with various dispersion. Note that there are positively strong linear relationship between Sales00 and Sales01, and between Inven00 and Inven01, respectively. For both of Industry XXX1 and XXX2, the correlation coefficients between Sales00 and Sales01 are greater than 0.82 and the correlation coefficients between Inven00 and Inven01 are greater than 0.98 for each stratum.

Moreover, Sales00 and Inven00 have moderately positive linear patterns. For Industry XXX1, the correlation coefficients within each stratum range from 0.40 to 0.98 except for stratum 4 in which the correlation coefficient is -0.04. For Industry XXX2, the correlation coefficients within each stratum range from 0.15 to 0.71. Thus, we should consider the correlation between Sales00 and Inven00 into our model so that the correlation structure can be maintained. Also we note that the range of each variable and pairwise correlations of other variables except (Sales01 versus Sales00) and (Inven01 versus Inven00) vary across strata, suggesting model heterogeneity across strata. Figure 3 and Figure 4 show the box plots of the log transformed variables against stratum for Industry XXX1 and Industry XXX2, respectively. As seen from those plots, the values of the six variables tend to be larger as the number of stratum increases. This indicates that stratum has information about our data as a factor and thus we make use of this information on our subgroup analysis.

2.2 Multivariate Regression Imputation

As mentioned in Section 2.1, there are distinct linear mean structures of the two study variables with other covariates and the study variables are also positively correlated. In order to further investigate data, it can be a good start to fit a multivariate linear regression model. We use the log-transformed data for all variables to have their distribution less skewed. For each industry, let $y_{j(s)}$ ($y_{j(i)}$) be the log-transformed current month sales (inventories) for jth unit ($j = 1, \ldots, n$). And let $x_{1j(s)}(x_{1j(i)})$ be the log-transformed prior month sales (inventories) and $x_{2j(s)}(x_{2j(i)})$ be the log-transformed administrative data value for sales (inventories).

Figure 5 and Figure 6 show four residual plots of the two study variables against their fitted values from the model for Industry XXX1 and XXX2, respectively. For both Industry XXX1 and XXX2, they indicate that there migth be two or three different variation patterns. This may result from different range of variables. For example, it seems more sound that small size of companies would have more variable data with respect to either sales or inventories or both of them. Here, the size may be defined as the size of monthly sales and inventories. Since units have different range of variables across strata, we use both residual plots as an indicator of homogeneity and information of strata to divide into different groups and separately fit a model for each group.

For Industry XXX1, we divide data into two groups: Group 1 consists of units within stratum 1 and 2, and Group 2 consists of units within stratum 3, 4, 5, and 6. We may call Group 1 small group and Group 2 non-small group. In order to account for difference across strata, we assume same slopes for current month (sales, inventories) but different regression intercepts. To do so, we define dummy variables for jth unit within hth stratum ($h = 1, \dots, 6$) as follows.

$$I_{hj} = \begin{cases} 1 & \text{if } j\text{th unit is within } h\text{th stratum} \\ 0 & \text{otherwise} \end{cases}$$

For jth unit in Group 1, we assume a multivariate linear model as follows.

$$\begin{bmatrix} y_{j(s)} \\ y_{j(i)} \end{bmatrix} = \begin{bmatrix} \beta^1_{0(s)} & \beta^1_{1(s)} & \beta^1_{2(s)} & \beta^1_{3(s)} & \beta^1_{4(s)} & \beta^1_{5(s)} \\ \beta^1_{0(i)} & \beta^1_{1(i)} & \beta^1_{2(i)} & \beta^1_{3(i)} & \beta^1_{4(i)} & \beta^1_{5(i)} \end{bmatrix} \begin{bmatrix} 1 \\ I_{2j} \\ x_{1j(s)} \\ x_{2j(s)} \\ x_{1j(i)} \\ x_{2j(i)} \end{bmatrix} + \begin{bmatrix} \epsilon^1_{j(s)} \\ \epsilon^1_{j(i)} \end{bmatrix},$$

where the $\underset{\sim}{\epsilon}^1_j = (\epsilon^1_{j(s)}, \epsilon^1_{j(i)})'$ are independent and identically distributed random vector with mean $\underset{\sim}{0}$ and variance-covariance matrix Σ^1. Note that $I_{j2} = 0$ indicates the jth unit within stratum 1.

For kth unit in Group 2, we assume the following multivariate linear model,

$$\begin{bmatrix} y_{k(s)} \\ y_{k(i)} \end{bmatrix} = \begin{bmatrix} \beta^2_{0(s)} & \beta^2_{1(s)} & \beta^2_{2(s)} & \beta^2_{3(s)} & \beta^2_{4(s)} & \beta^2_{5(s)} & \beta^2_{6(s)} & \beta^2_{7(s)} \\ \beta^2_{0(i)} & \beta^2_{1(i)} & \beta^2_{2(i)} & \beta^2_{3(i)} & \beta^2_{4(i)} & \beta^2_{5(i)} & \beta^2_{6(s)} & \beta^2_{7(s)} \end{bmatrix} \begin{bmatrix} 1 \\ I_{4k} \\ I_{5k} \\ I_{6k} \\ x_{1k(s)} \\ x_{2k(s)} \\ x_{1k(i)} \\ x_{2k(i)} \end{bmatrix} + \begin{bmatrix} \epsilon^2_{k(s)} \\ \epsilon^2_{k(i)} \end{bmatrix},$$

where the $\underset{\sim}{\epsilon}^2_k = (\epsilon^2_{k(s)}, \epsilon^2_{k(i)})'$ are independent and identically distributed random vector with mean $\underset{\sim}{0}$ and variance-covariance matrix Σ^2. Note that $(I_{4k}, I_{5k}, I_{6k}) = (0, 0, 0)$ indicates the kth unit within stratum 3. In this study, we only assume the homoegeneity of variance within each group for both Industry XXX1 and XXX2 since given data are strongly right-skewed so that it fails to assume the normality.

Suppose that due to item nonresponse, only $n_1 (< n)$ of the $\underset{\sim}{y} = (y_{(s)}, y_{(i)})'$ values are observed with $n_0 = n - n_1$ missing and the missing mechanism is ignorable. We consider only one missing pattern where both $y_{(s)}$ and $y_{(i)}$ are missing. Let $A_{mis} = \{j \in A : \underset{\sim}{y}_j$ is missing$\}$ and $A_{obs} = \{j \in A : \underset{\sim}{y}_j$ is observed$\}$, where A is the set of indices j in the data. For each $j \in A_{mis}$, the deterministic regression imputation uses its predicted value as an imputed value. More specifically, let G_g denote the set of indices which belong to Group $g(g = 1, 2)$ and let $\underset{\sim}{x}'_j = (1, x_{1j(s)}, x_{2j(s)}, x_{1j(i)}, x_{2j(i)})$ be its corresponding covariates which are always observed. Then, we can use $\hat{\underset{\sim}{\beta}}^{g\prime}\underset{\sim}{x}_j$ as an imputed value, say $\underset{\sim}{z}_j$, where $\hat{\underset{\sim}{\beta}}^g$ is the ordinary least square estimate of $\underset{\sim}{\beta}^g$ and $j \in A_{mis} \cap G_g$.

2.3 The Shao-Sitter Bootstrap Variance Estimation

For variance estimation after imputation, we apply the bootstrap procedure proposed by Shao and Sitter(1996). In this section, we describe how to obtain the bootstrap estimator in detail. Note that the simulated data sets are stratified simple random sample without replacement, as mentioned above. In order to mimic the without replacement nature of the original sampling design, we use the without-replacement bootstrap (BWO) proposed by Sitter(1992).

We use the following three-step algorithm proposed by Shao and Sitter(1996). Let h, n_h and $H(= 6)$ denote hth stratum, the number of units within the hth stratum and the number of strata, respectively.

$Step$ 1) Draw a bootstrap sample, say $\{\underset{\sim}{y}^*_{hj}\}$, using the method of BWO from the imputed sample $Y^I = \{\underset{\sim}{y}^I_{hj} : j = 1, \ldots, n_h\}, h = 1, \ldots, H$, independently across strata, where $\underset{\sim}{y}^I_{hj} = \underset{\sim}{y}_{hj}$ if $j \in A_{obs}$ and $\underset{\sim}{y}^I_{hj} = \underset{\sim}{z}_{hj}$ if $j \in A_{mis}$, $h = 1, \ldots, H$.

$Step$ 2) For jth unit within hth stratum, let a^*_{hj} be the response indicator associated with $\underset{\sim}{y}^*_{hj}$, $A^*_{mis} = \{(h, j) : a^*_{hj} = 0\}$, $A^*_{obs} = \{(h, j) : a^*_{hj} = 1\}$, $Y^*_{mis} = \{\underset{\sim}{y}^*_{hj} : (h, j) \in A^*_{mis}\}$ and $Y^*_{obs} = \{\underset{\sim}{y}^*_{hj} : (h, j) \in A^*_{obs}\}$. Apply the same imputation procedure used in constructing the original imputed data set to the units in A^*_{mis} using Y^*_{obs}. Denote Y^{*I} the boostrap analog of Y^I.

$Step$ 3) Repeat $Step$ 1 to 3 to draw B bootstrap samples from Y^I.

For our interest, say θ, and its estimator, say $\hat{\theta}$, obtain the bootstrap analog $\hat{\theta}^*$ of $\hat{\theta}$, based on the imputed bootstrap data set Y^{*I}. Then we obtain the following variance estimator by using the Monte Carlo approximation,

$$\frac{1}{B}\sum_{b=1}^{B}(\hat{\theta}^{*b} - \bar{\theta}^*)^2,$$

where $\hat{\theta}^{*b} = \hat{\theta}^{*b}(Y^{*b})$, $b = 1, \ldots, B$ and $\bar{\theta}^* = B^{-1}\sum_{b=1}^{B}\hat{\theta}^{*b}$. Here, Y^{*b} are bth independent bootstrap sample. Note that this bootstrap procedure imputes the bootstrap data sets in exactly the same way that the original data set is imputed in $Step$ 2.

We also use the BWO method discussed by Sitter(1992) in order to generate bootstrap samples as follows. Under stratified simple random sampling without replacement, let $n'_h = n_h - (1 - f_h)$ and $k_h = \frac{N_h}{n_h}(1 - \frac{(1-f_h)}{n_h})$ where N_h is the number of units within hth stratum of finite population and $f_h = n_h/N_h$ for $h = 1, \ldots, H$. Then, the algorithm is as follows.

$Step$ 1) Create the hth stratum of the pseudopopulation by replicating $\{\underset{\sim}{y}_{hj}\}_{j=1}^{n_h}$ k_h times. Repeat this for each stratum.

$Step$ 2) Resample n'_hunits from stratum h without replacement to get $\{\underset{\sim}{y}^*_{hj}\}_{j=1}^{n'_h}$ for $h = 1, \ldots, H$.

Here, n'_h and k_h are chosen to satisfy the following: $f^*_h = f_h$ and

$$Var^*(\bar{\underset{\sim}{y_h}}^*) \quad = \quad \frac{(1 - f_h}{n_h(n_h - 1)} \sum_{j=1}^{n_h} (\underset{\sim}{y}_{hj} - \bar{\underset{\sim}{y}}_h)(\underset{\sim}{y}_{hj} - \bar{\underset{\sim}{y}}_h)',$$

where $f^*_h = n'_h/(k_h n_h)$ is the resampling fraction.

This algorithm only makes sense if n'_h and k_h are both integers for all $h = 1, \ldots, H$ and clearly since $0 \leq f_h \leq 1$, $n'_h$is noninteger unless $f_h = 0$ or 1. To avoid this problem, one randomizes bewteen bracketing integer values, discussed by Sitter(1992).

3. Results

In this section, we present several results from application of the two methods described in Section 2. Table 3(Table 6) shows the estimated coefficients and variance-covariance matrices by each group for Industry XXX1(XXX2). As expected from the exploratory data analysis, the variance-covariance matrix for Group 1(Small) has a higher values than that for Group 2(Non-small) and regression coefficients also have different values between two groups. However, someone may be reluctant to have some negative coefficients of the covariates because it would be a bit different from our intuition. This may result from multicollinearity problem. Note that our interest is to estimate the population total of each study variable, not to estimate and interpret the regression coefficient parameters. In this case, multicollinearity seldom changes the esimated total values and thus we do not deal with it in this study.

In Figure 7 and 8, several residual plots are presented for Group 1 and 2, respectively. Compared to the plots from the exploratory data analysis, they support the assumption of homogeniety, showing random scatters.

Table 4 presents comparison of the true population values and regression imputation estimates for Industry XXX1. They look similar but regression imputation estimates seem to be underestimated. The estimated totals for Sales00 and Inven00 are 47,894,131,580 and 99,778,621,944, respectively. As shown in Table 5, the standard errors of the estimated totals of Sales00 and Inven00 are $483, 102, 506$ and $1, 743, 125, 258$, respectively and their asymptotic 95% confidence intervals contain the true populaion values.

Almost the same imputation approach is applied to Industry XXX2. Based on the exploratory data analysis, we divide data into three groups: Group 1(Small) consists of units within stratum 1, Group 2(Medium) consists of units within stratum 2, 3, 4 and 5, and Group 3(Large) consists of units within stratum 6. Similarly, we assume same slopes for current month (sales, inventories) but different regression intercepts. The estimated coefficients and variance-covariance matrix for each group for Industry XXX2 are present in Table 6. It seems that Group 1 and 2 have similar estimated slopes of the four covariates but different estimated variance-covariance matrices. Especially, there is a little large difference in variance of Sales00 between Group 1 and 2. Group 3 has different estimated slopes of the four covariates as well as different estimated variance-covariance matrix from those of Group 1 and 2. In Figures 9, 10 and 11, residual plots are presented for Group1, Group2 and Group 3, respectively. They support the model assumption which is the homogeneous variance structure within each group. Table 7 shows comparison of the true population values and regression imputation estimates for Industry XXX2. Also, the bootstrap variance estimates and their asymptotic 95% confidence intervals for totals of Sales00 and Inven00 are present in Table 8. Similarly, the asymptotic confidence intervals also contain the true populaion values.

4. Discussion

We study the two simulated data sets which have distinct linear mean structures of two study variables with other covariates. The two study variables which are subject to missingness are positively correlated. In order to preserve the correlation and reflect the linear mean structures of the two study variables, we assume multivariate linear regression models. In order to assume homogeneity of variance structure, we divide data into sub-groups and fit a different model for each group. Since data have strongly right-skewed distribution, we do not assume normality. Based on this model, we impute missing values using the deterministic regression imputation and then use the Shao-Sitter bootstrap procedure for variance estimation. For both of the two simulated data sets, the regression imputation estimates of totals for Sales00 and inven00 are provided and their asymptotic 95% confidence intervals, which are based on the consistency of the bootstrap variance estimator, contain the true population values.

REFERENCES

Bickel, P. J. and Freedman, D. A. (1984), "Asymptotic Normality and the Bootstrap in Stratified Sampling," *The Annals of Statistics*, 12, 470-482.

Efron, B. (1994), "Missing Data, Imputation, and the Bootstrap," *Journal of the American Statistical Association*, 89, 463-479.

Enders, Craig. (2010), "Applied Missing Data Analysis," *Guilford Press: New York*.

Fay, R. E. (1993), "Valid Inferences From Imputed Survey Data," *in Proceedings of the Survey Research Methods Section, American Statistical Association*, 41-48.

Kim, J. K. and Shao, J. (2013), "Statistical Methods for Handling Incomplete Data," *CRC Press Talyor & Francis Group*.

Little, R. J. and Rubin, D. B. (2002), "Statistical Analysis with Missing Data (2nd Edition)," *John Wiley & Sons, Hoboken, NJ*.

Schafer, J. L. (1997), "Analysis of Incomplete Multivariate Data (First Edition)," *Chapman & Hall, London*.

Shao, J. and Sitter, R. R. (1996), "Bootstrap for Imputed Survey Data," *Journal of the American Statistical Associations*, 91(435), 1278-1288.

Sitter, R. R. (1992), "Comparing Three Bootstrap Methods for Survey Data," *The Canadian Journal of Statistics*, 20(2), 135-154.

Tanner, M. A. (1993), "Tools for Statistical Inference, Methods for the Exploration of Posterior Distributions and Likelihood Functions (Second Edition)," *Springer-Verlag New York, Inc*.

Table 1: Summary statistics of the six variables for Industry XXX1

Variables	Min.	Median	Mean	Max
Sales00	27651	5037715	17002414	572063107
Inven00	56888	9502698	29329187	887229044
Sales01	1455	3194750	13319990	585251680
Inven01	13038	6638809	23101331	824741096
Asales00	1555	2697752	13145460	572137959
Ainven00	14781	7920970	24758649	887126854

NOTE: For Sales00 and Inven00 which are subject to missingness, the summary statistics are only for respondents.

Table 2: Summary statistics of the six variables for Industry XXX2

Variables	Min.	Median	Mean	Max
Sales00	4702	277203	1159863	48149601
Inven00	992	320165	1811902	101653240
Sales01	2916	178037	927408	50354693
Inven01	258	179137	1421635	107049635
Asales00	2573	163237	910110	48234335
Ainven00	-318	219949	1449015	101380478

NOTE: For Sales00 and Inven00 which are subject to missingness, the summary statistics are only for respondents.

Table 3: Multivariate regression imputation point estimates for Industry XXX1

Coefficients	Small (l.Sales00, l.Inven00)	Non-small (l.Sales00, l.Inven00)
Intercept	(0.067, -0.198)	(0.026, -0.292)
Stratum 2	(-0.013, -0.010)	-
Stratum 4	-	(0.002, 0.001)
Stratum 5	-	(-0.001, -0.002)
Stratum 6	-	(-0.030, -0.271)
l.Sales01	(0.592, 0.590)	(0.893, 0.895)
l.Asales00	(0.411, -0.588)	(0.101, -0.901)
l.Inven01	(0.410, 0.411)	(0.101, 0.099)
l.Ainven00	(-0.409, 0.589)	(-0.095, 0.907)
Variance- Covariance Matrix	0.0007 0.0007 0.0007 0.0007	0.0003 0.0003 0.0003 0.0003

NOTE: These are the estimates obtained from the multivariate linear model for log-transformed variables(log(Sales00), log(Inven00), log(Sales01), log(Inven01), log(Asales00) and log(Ainven00)). Stratum 2, 4, 5 and 6 are the indicator variables which have a value of 1 for being within the corresponding stratum, otherwise, have a value of 0.

Table 4: Comparison of the true population values and estimates for Industry XXX1

		True	Estimate
Sales00	Total	4.83E+10	4.79E+10
	Mean	2,304,717	2,283,174
	Variance	4.38E+13	4.37E+13
	Skewness	49.67	49.71
Inven00	Total	10.09E+10	9.98E+10
	Mean	4,807,697	4,756,573
	Variance	1.19E+14	1.17E+14
	Skewness	39.02	40.03
Sales00 and Inven00	Correlation	0.97	0.97

Table 5: Bootstrap variance estimates and 95% confidence intervals for each total of Sales00 and Inven00 for Industry XXX1

	Sales00	Inven00
Population	48,346,053,043	100,851,062,160
Estimate	47,894,131,580	99,778,621,944
s.e	483,102,506	1,743,125,258
95% CI	(46,947,250,669, 48,841,012,492)	(96,362,096,439, 103,195,147,449)

Table 6: Multivariate regression imputation point estimates for Industry XXX2

Coefficients	Small (l.Sales00, l.Inven00)	Medium (l.Sales00, l.Inven00)	Large (l.Sales00, l.Inven00)
Intercept	(0.401, 0.075)	(0.435, 0.039)	(0.016, 0.034)
Stratum 3	-	(0.004, 0.003)	-
Stratum 4	-	(0.023, 0.016)	-
Stratum 5	-	(0.025, 0.013)	-
l.Sales01	(0.052, -0.016)	(0.015, 0.017)	(0.475, 0.009)
l.Asales00	(0.923, 0.005)	(0.976, -0.021)	(0.522, -0.010)
l.Inven01	(0.929, 1.003)	(0.963, 0.972)	(-0.043, 0.950)
l.Ainven00	(-0.916, 0.0004)	(-0.967, 0.028)	(0.045, 0.049)
Variance-	0.0034 0.0003	0.0006 0.0005	0.0016 -0.0001
Covariance Matrix	0.0003 0.0004	0.0005 0.0005	-0.0001 0.0004

NOTE: These are the estimates obtained from the multivariate linear model for log-transformed variables(log(Sales00), log(Inven00), log(Sales01), log(Inven01), log(Asales00) and log(Ainven00)). Stratum 3, 4, and 5 are the indicator variables which have a value of 1 for being within the corresponding stratum, otherwise, have a value of 0.

Table 7: Comparison of the true population values and estimates for Industry XXX2

		True	Estimate
Sales00	Total	1.68E+09	1.68E+09
	Mean	118,988.36	119,344.90
	Variance	2.93E+11	2.92E+11
	Skewness	55.94	56.27
Inven00	Total	1.98E+09	2.10E+09
	Mean	140,538.2	149,059.1
	Variance	1.63E+12	1.62E+12
	Skewness	63.82	63.92
Sales00 and Inven00	Correlation	0.75	0.75

Table 8: Bootstrap variance estimates and 95% confidence intervals for each total of Sales00 and Inven00 for Industry XXX2

	Sales00	Inven00
Population	1,677,378,977	1,981,167,030
Estimate	1,682,405,706	2,101,285,610
s.e.	35,478,111	70,078,756
95% CI	(1,612,868,608, 1,751,942,804)	(1,963,931,249, 2,238,639,972)

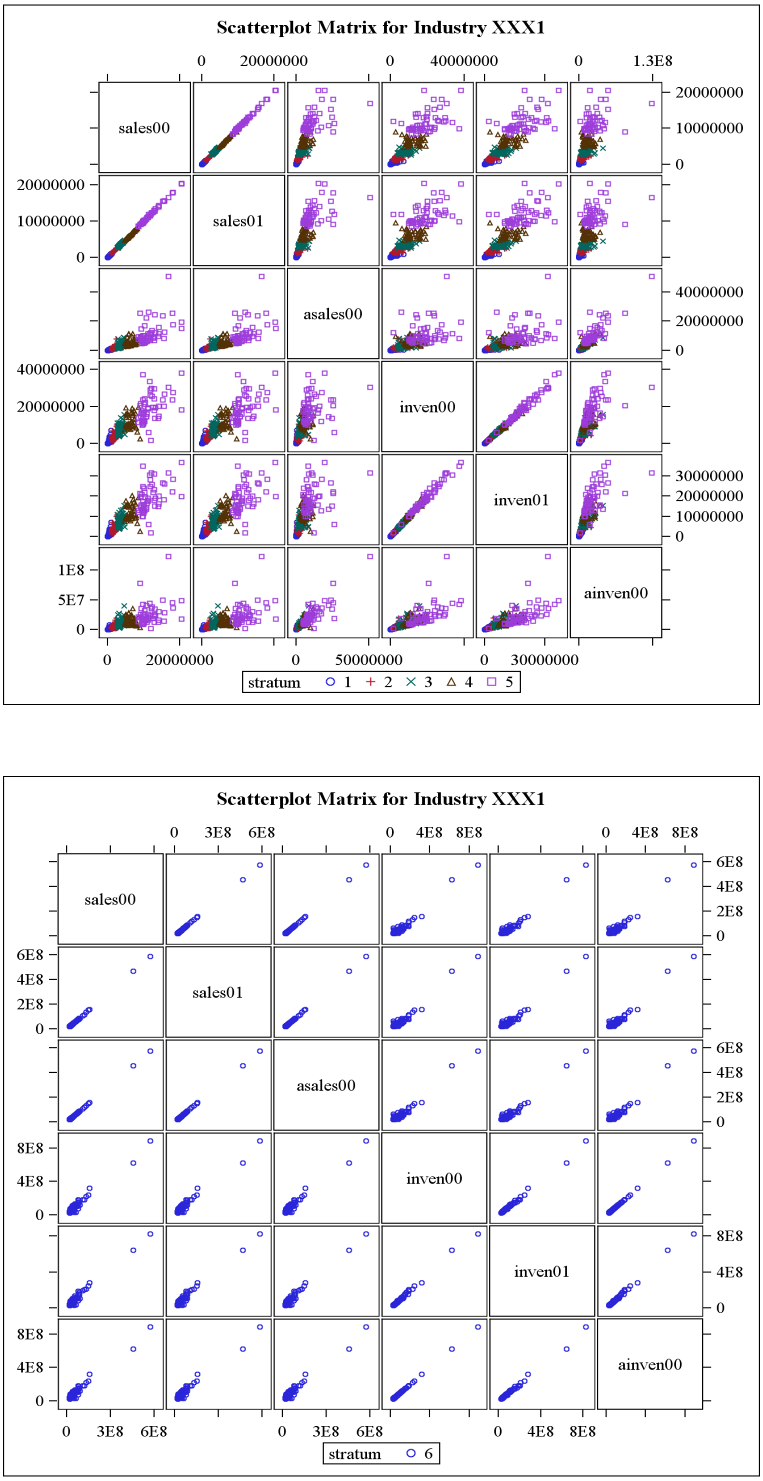
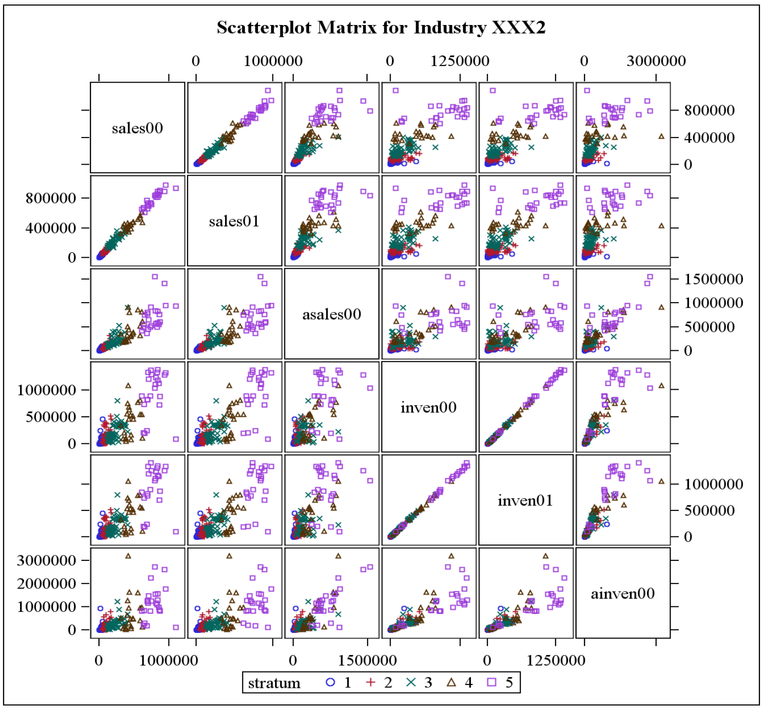


Figure 1: Scatter plot matrix for Industry XXX1 (The upper is the scatter plot matrix for first five strata and the bottom is for stratum 6.)



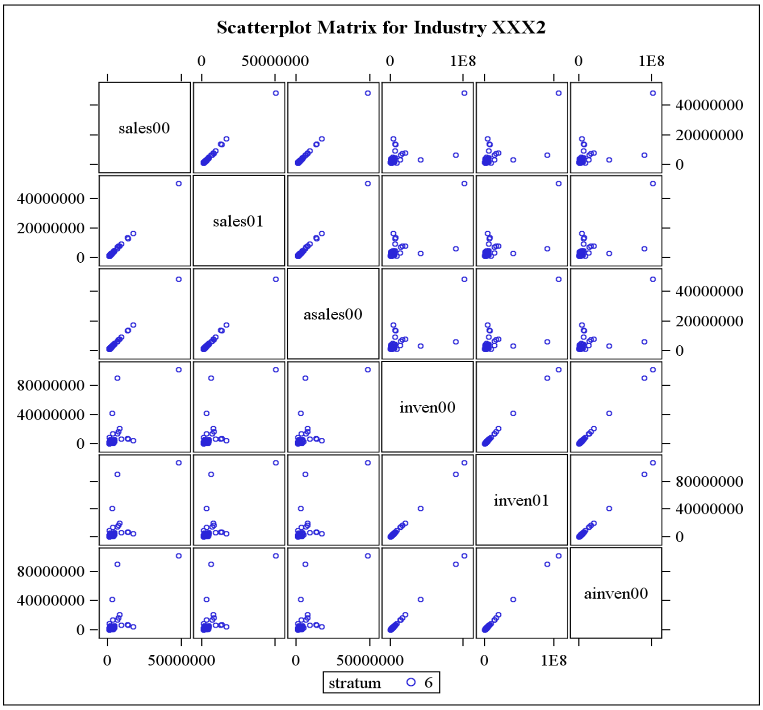


Figure 2: Scatter plot matrix for Industry XXX2 (The upper is the scatter plot matrix for first five strata and the bottom is for stratum 6.)

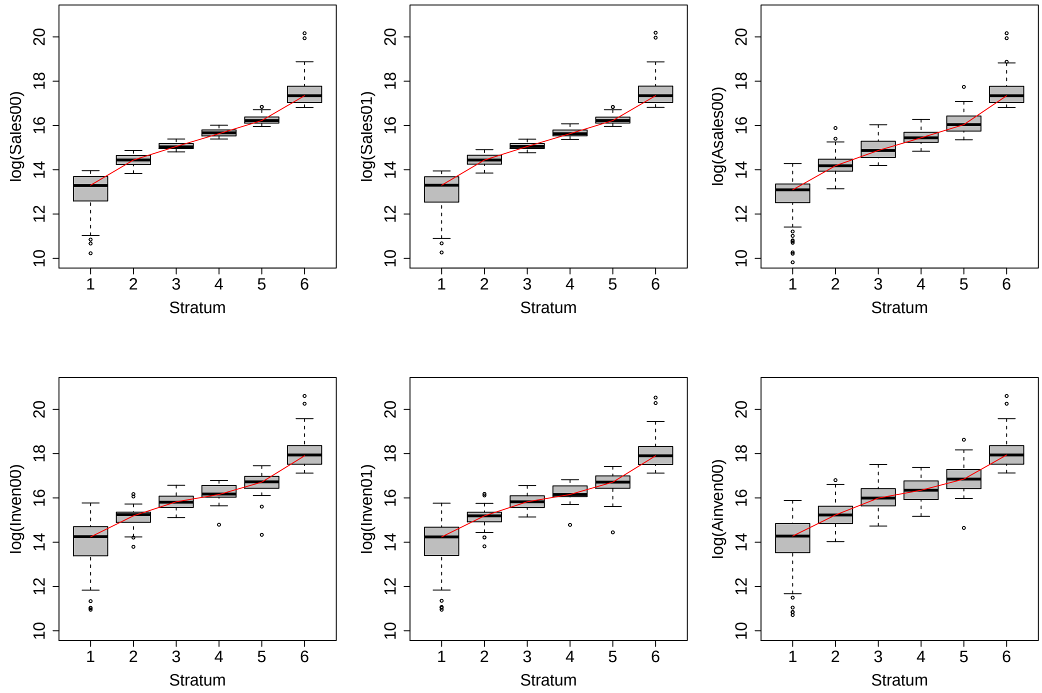


Figure 3: Box plots of the six log-transformed variables by stratum for Industry XXX1

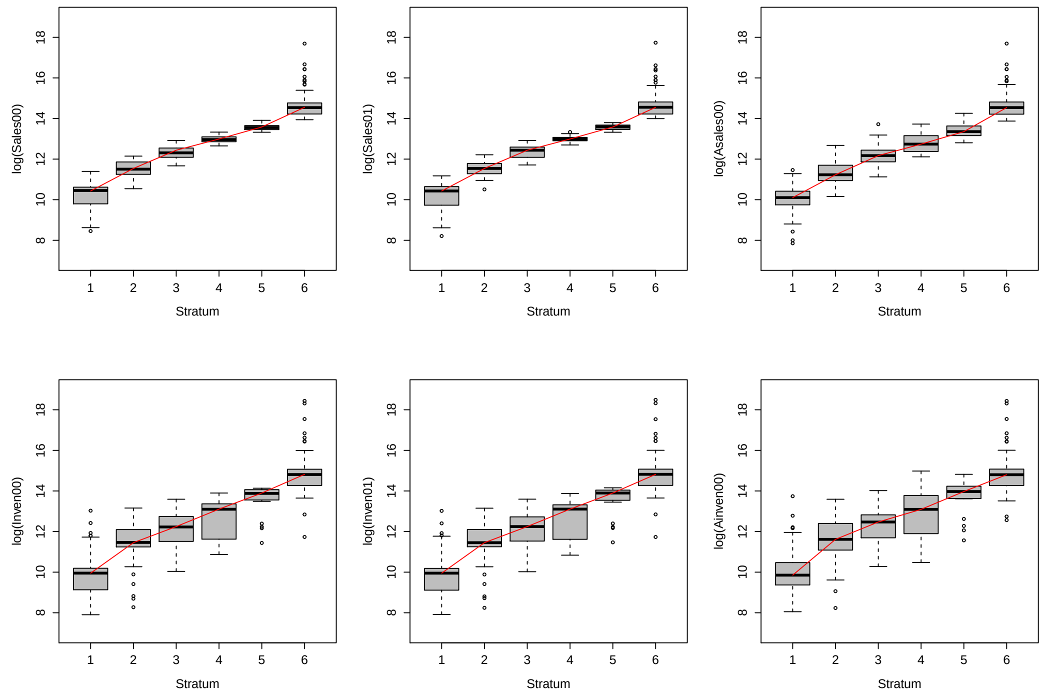


Figure 4: Box plots of the six log-transformed variables by stratum for Industry XXX2

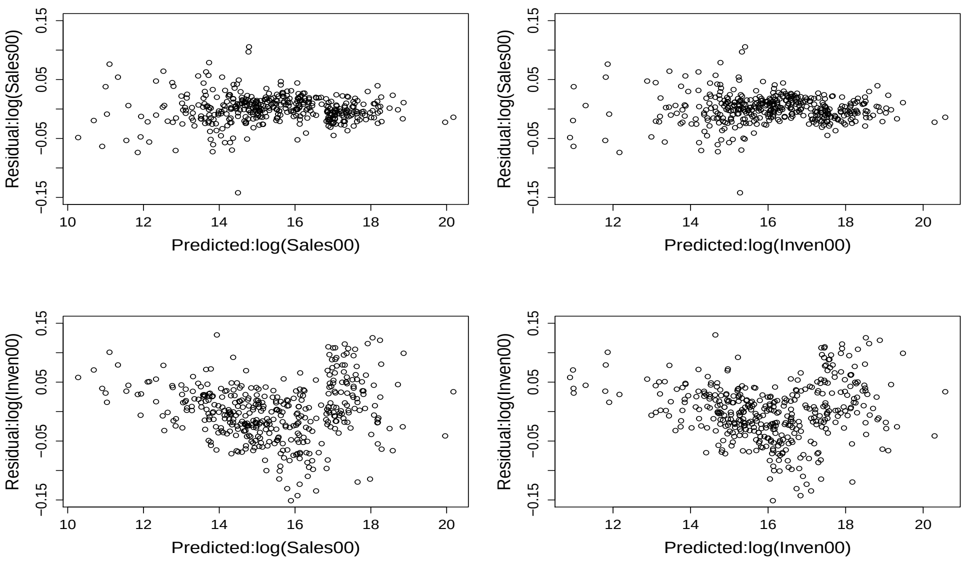


Figure 5: Residual plots of the log transformed responses against predicted values for all strata (Industry XXX1)

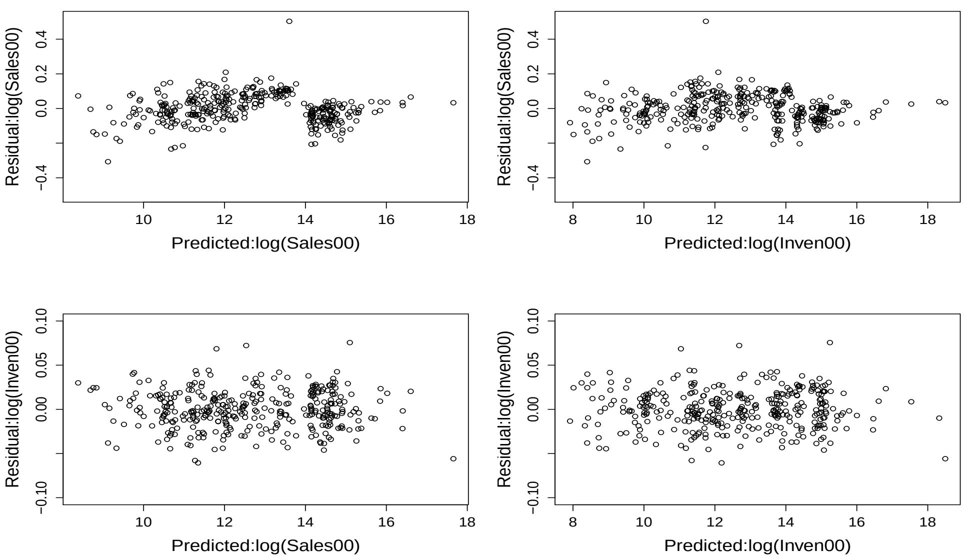


Figure 6: Residual plots of the log transformed responses against predicted values for all strata (Industry XXX2)

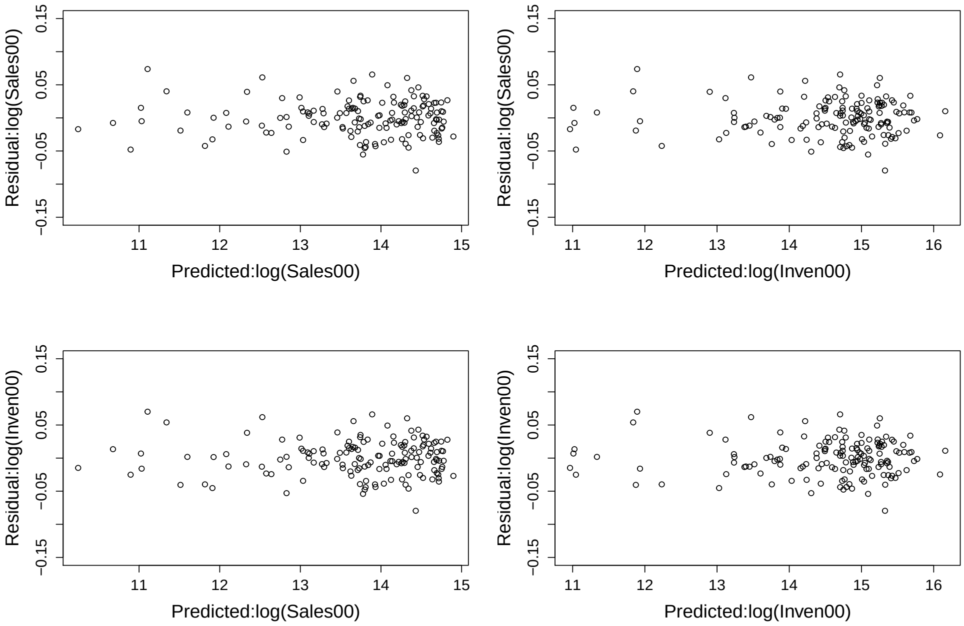


Figure 7: Residual plots of the log transformed responses against predicted values for Group 1 (Industry XXX1)

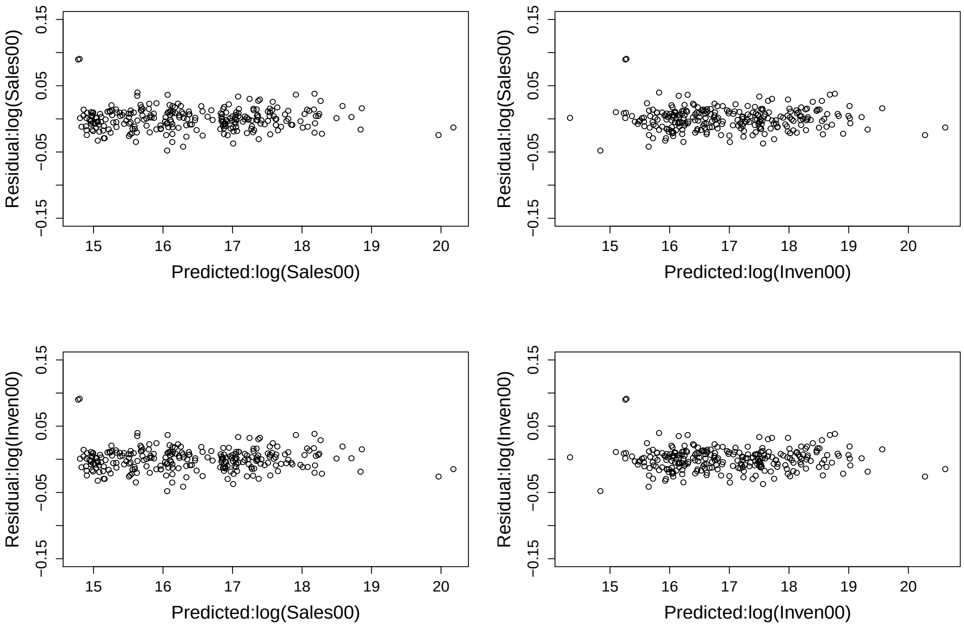


Figure 8: Residual plots of the log transformed responses against predicted values for Group 2 (Industry XXX1)

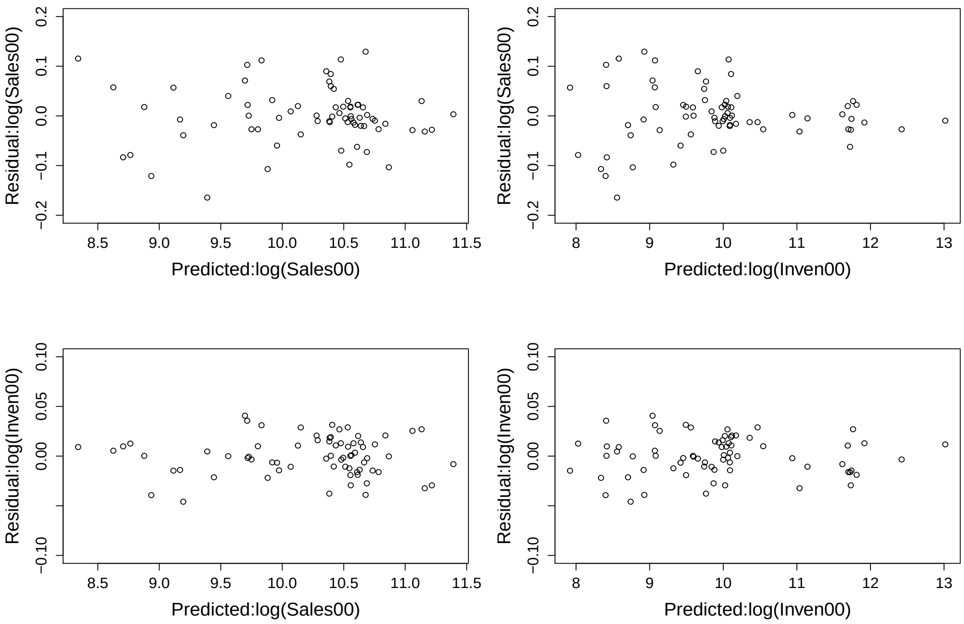


Figure 9: Residual plots of the log transformed responses against predicted values for Group 1 (Industry XXX2)

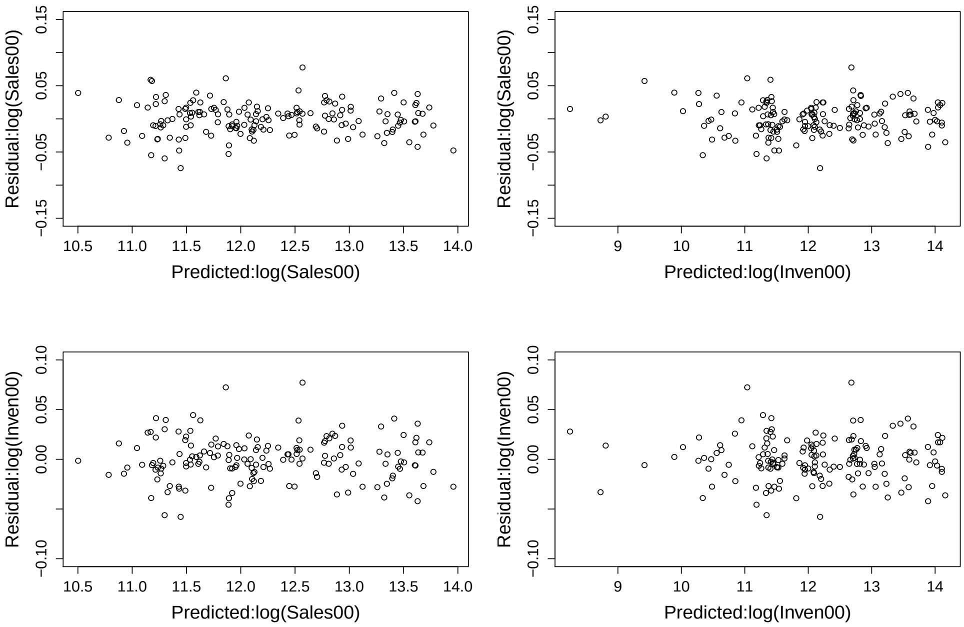


Figure 10: Residual plots of the log transformed responses against predicted values for Group 2 (Industry XXX2)

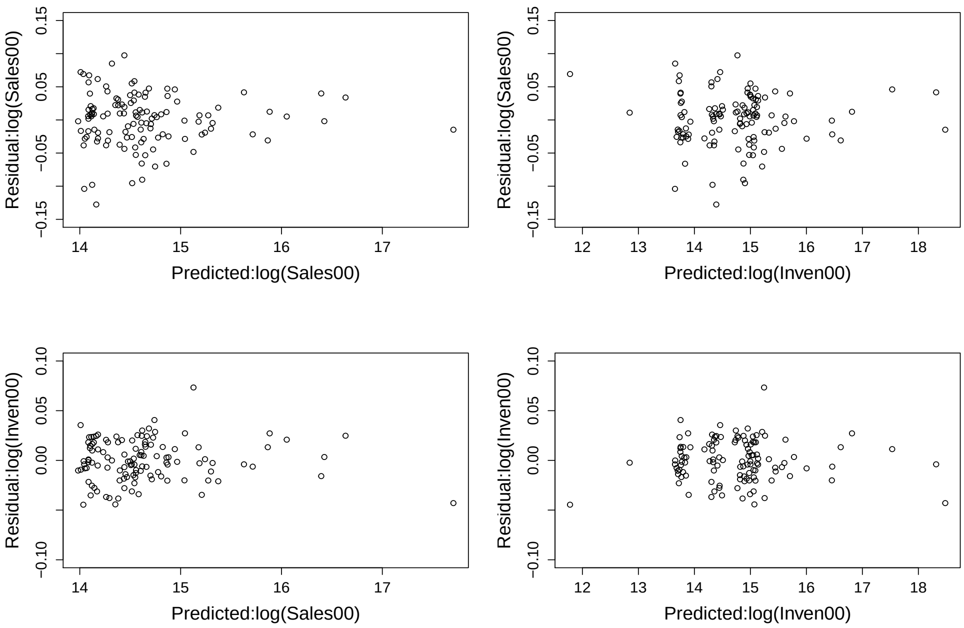


Figure 11: Residual plots of the log transformed responses against predicted values for Group 3 (Industry XXX2)