Paper 370-2008

On Predicting Log-transformed Linear Models with Heteroscedasticity

Siwei Jia, Momentum Market Intelligence, Portland, Oregon
Sarika Rathi, Momentum Market Intelligence, Portland, Oregon

ABSTRACT

For log-transformed linear model $\mathbf{E}(\log \mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$, one of the most common practices is to predict at the original scale of dependent variable $\mathbf{Y}$. It is well known that a simple exponentiation $\exp(\mathbf{X}\hat{\boldsymbol{\beta}})$ will cause bias, even in the case when $\hat{\boldsymbol{\beta}}$ is $BLUE$. The bias is worse if heteroscedasticity is associated with the log-transformed model. In this paper we propose an anti-log transformation that is practically simple yet theoretically sound. We use Base SAS® and SAS/STAT® to transform, fit and adjust the predictions. This paper is intended for statisticians, econometricians and other quantitative analysts concerned with application of log-transformed models.

Keywords: log-transform, multiplicative model, heteroscedasticity, weighted least square, prediction, SAS GLM procedure.

INTRODUCTION

We consider the following log-transformed linear model:

$$\log Y_{ij} = \mu_i + \beta_1 X_{1ij} + \cdots + \beta_p X_{pij} + \varepsilon_{ij}, \tag{1}$$

where Y_{ij} is the dependent variable for the jth observation in ith group; μ_i is a constant associated with the ith group; $X_{1ij}, \cdots, X_{pij}$ are the p covariates in model (1). We assume that the error term $\varepsilon_{ij} \sim Normal(0, \sigma_i^2)$ and ε_{ij} and ε_{st} are independent for all $i \neq s$ or $j \neq t$. Notice that many multiplicative models can be transformed into the form of model (1). Model (1) could also be applied to cross-sectional time series if the above variance structure were accepted as a good approximation.

When $\sigma_i^2 = \sigma^2 \ \forall i$, we have

$$\mathbf{E}Y_{ij} = \exp(\mu_i + \sum_{k=1}^{p} \beta_k X_{kij} + \frac{1}{2}\sigma^2).$$

Otherwise model (1) is heteroscedastic and we may use the weighted least square approach, i.e.,

$$\frac{\log Y_{ij}}{w_i} = \mu_i^* \frac{1}{w_i} + \beta_1^* \frac{X_{1ij}}{w_i} + \cdots + \beta_p^* \frac{X_{pij}}{w_i} + \frac{\varepsilon_{ij}}{w_i}, \tag{2}$$

in which case if σ_i is chosen for w_i then

$$\mathbf{E}Y_{ij} = \exp(\mu_i^* + \sum_{k=1}^{p} \beta_k^* X_{kij} + \frac{1}{2}\sigma_i^2). \tag{3}$$

In practice, however, we usually can only estimate μ_i^* and β_k^* without knowing σ_i^2. Literatures can be found regarding estimating μ_i^*, β_k^* and σ_i^2 (e.g., Teedens and Koerts 1972). It is beyond the scope of this paper to discuss the estimation issues in detail. Focusing on transformation, we suppose that consistent $BLUE$ $\hat{\mu_i^*}$, $\hat{\beta_k^*}$ and $\hat{\sigma_i^2}$ are obtained. Then

$$\widehat{Y_{ij}} = \exp(\hat{\mu_i^*} + \sum_{k=1}^{p} \hat{\beta_k^*} X_{kij} + \frac{1}{2}\hat{\sigma_i^2}) \tag{4}$$

is also a consistent estimate. The estimate $\widehat{Y_{ij}}$ is widely used for prediction (e.g., Miller 1984). However, estimate $\widehat{Y_{ij}}$ given by equation (4) over-estimates the expected value in equation (3) because of convexity of exponential function

(Jensen's Inequality). In general, making adjustment to $\widehat{Y_{ij}}$ to achieve unbiasedness is computationally expensive. However, we can have a simpler adjustment at a relative low cost. Notice that $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\beta}}$ are jointly normal given the assumptions of models (1) and (2). We ignore correlations between $\hat{\mu_i^*}$'s, $\hat{\beta_k^*}$'s and $\hat{\sigma_i^2}$'s, and write equation (4) as

$$\widehat{Y_{ij}} = e^{\hat{\mu_i^*}} \left(\prod_{k=1}^{p} e^{\hat{\beta_k^*} X_{kij}} \right) e^{\frac{1}{2}\hat{\sigma_i^2}}.$$

Since

$$\mathbf{E}\exp(\hat{\mu_i^*}) = \exp(\mu_i^* + \frac{1}{2}\sigma^2_{\hat{\mu_i^*}})$$

and

$$\mathbf{E}\exp(\hat{\beta_k^*} X_{kij}) = \exp(\beta_k^* X_{kij} + \frac{1}{2}(X_{kij}\sigma_{\hat{\beta_k^*}})^2),$$

we have

$$\widetilde{Y_{ij}} = \exp(\hat{\mu_i^*} - \frac{1}{2}\sigma^2_{\hat{\mu_i^*}}) \left[\prod_{k=1}^{p} \exp(\hat{\beta_k^*} X_{kij} - \frac{1}{2}(X_{kij}\sigma_{\hat{\beta_k^*}})^2) \right] \exp(\frac{1}{2}\hat{\sigma_i^2}). \tag{5}$$

For simplicity we omit the adjustment for $\exp(\frac{1}{2}\hat{\sigma_i^2})$. It turns out that estimate $\widetilde{Y_{ij}}$ given by equation (5) significantly reduces bias and it is relatively easy to compute by using SAS[1]. In the next section we give details of fittings and adjustments by using SAS, which is outlined below:

1. fit model (1) by *PROC GLM*; discuss problems caused by anti-log and heteroscedasticity;
2. save the residuals from the above fitting; test heteroscedasticity; and obtain weights w_i for model (2);
3. fit model (2) to get $\hat{\mu_i^*}$, $\widehat{\sigma_{\hat{\mu_i^*}}}$, $\hat{\beta_k^*}$, $\widehat{\sigma_{\hat{\beta_k^*}}}$ and $\hat{\sigma_i^2}$ $(= w_i^2 MSE)$;
4. compute $\widetilde{Y_{ij}}$ based on equation (5) by setting $\sigma_{\hat{\mu_i^*}} = \widehat{\sigma_{\hat{\mu_i^*}}}$ and $\sigma_{\hat{\beta_k^*}} = \widehat{\sigma_{\hat{\beta_k^*}}}$.

We also provide some simulation results in the next section.

ESTIMATION AND PREDICTION

For this study we simulate a simple marketing research data. We assume that sales of a commodity depend on its price level, promotion and group. A balanced data is generated with 300 observations, i.e., 30 observations for each of the 10 groups with different group means. Covariate X_1 is a dummy variable indicating presence of a promotion; X_2 is logarithm of price level. The error term is generated such that variance differs among the ten groups. Finally, dependent variable $\log Y$, the logarithm of sales, is generated according to model (1) for $p = 2$. The following is the scatter plot of the generated dataset.

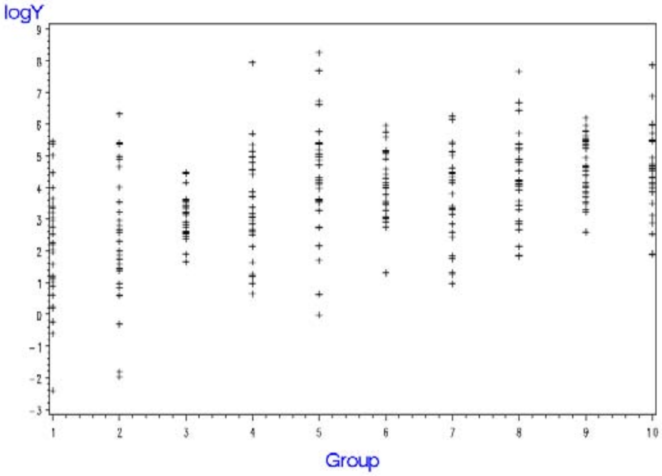


Figure 1: Scatter Plot of $\log Y$ against Group

[1] More comprehensive adjustment is possible but it adds complexity in practice.

We first make an OLS fitting by SAS/STAT GLM procedure.

```
PROC GLM DATA = model_dataset OUTSTAT = output_anova1;
    CLASS Group;
    MODEL logY = Group X1 X2 / NOINT SOLUTION;
    OUTPUT OUT = out1 RESIDUAL = ols_residual PREDICTED = y_forecast1;
    ODS OUTPUT PARAMETERESTIMATES = output_estimate1;
RUN; QUIT;
```

A naïve way to predict Y is simply taking exponential of *y_forecast1* from the above GLM model. As shown in the plot of group means (Figure 2), it leads to serious under-estimation since bias is associated with anti-log transformation.

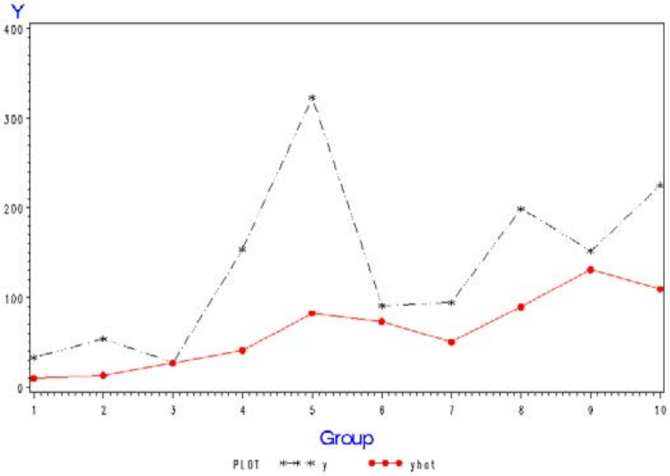


Figure 2: Naïve Anti-Log without Adjustment

As suggested in Section 1 we make adjustments based on equations (4) and (5) to remove bias associated with anti-log transformation. Though these adjustments improve predictions still it is not satisfactory, as illustrated by plotting group means in Figure 3.

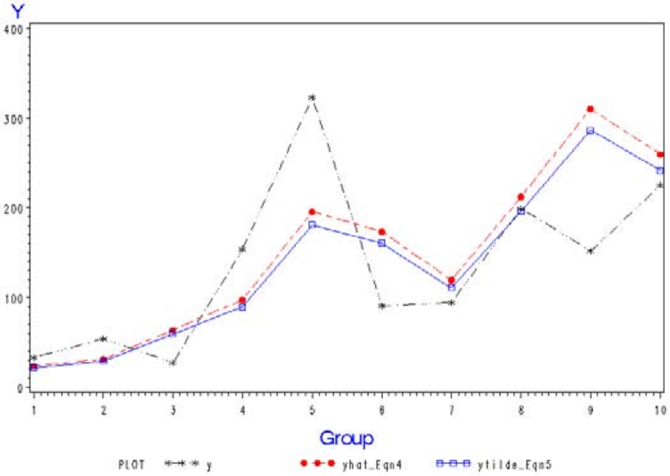


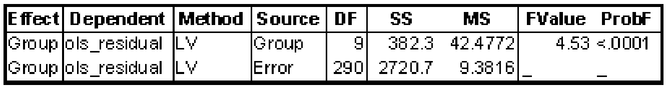
Figure 3: $\hat{Y}$ and $\tilde{Y}$ from OLS

This is because in the OLS fitting σ_i^2's are not properly estimated due to heteroscedasticity. To see this we saved the residuals from the above OLS fitting and run Levene's test, which is conveniently provided by SAS to test heteroscedasticity in the fitted model.

```
PROC GLM DATA = out1;
    CLASS Group;
    MODEL ols_residual = Group;
    MEANS Group / HOVTEST;
```

```
        ODS OUTPUT HOVFTEST = Levene;
    RUN; QUIT;
```

Results in Table 1 indicate that the null hypothesis of homoscedasticity should be rejected.



Effect	Dependent	Method	Source	DF	SS	MS	FValue	ProbF
Group	ols_residual	LV	Group	9	382.3	42.4772	4.53	<.0001
Group	ols_residual	LV	Error	290	2720.7	9.3816	_	_

Table 1: Results from Levene's Test

In order to correct for heteroscedasticity we apply weighted least square (WLS) method using error standard deviation as weight (w_i in model (2)). However, care should be taken when the WEIGHT statement in GLM procedure is used. The normal equations for weighted GLM in SAS are in the following form (SAS Institute 2007):

$$\mathbf{X}'\mathbf{W}\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{W}\mathbf{Y},$$

where $\mathbf{W}$ is a diagonal matrix consisting of the values of the variable specified in the WEIGHT statement. Therefore we need to apply the *reciprocal of error variance* as weight for the WEIGHT statement in order to achieve results from model (2). The following is the SAS code to obtain the weights.

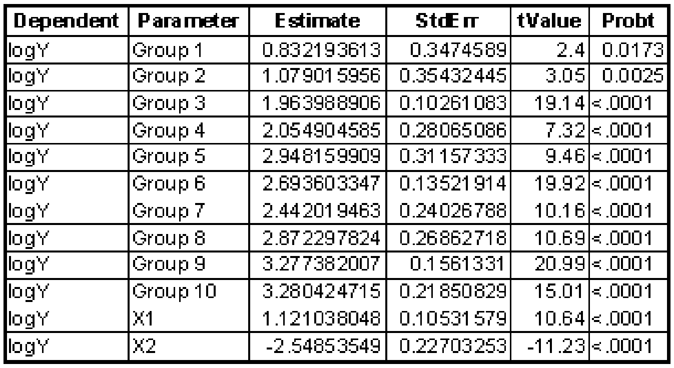
```
    PROC SQL NOPRINT;
        CREATE TABLE group_stderr AS
            SELECT Group, STD(ols_residual) AS err_std
            FROM out1
            GROUP BY Group;

        CREATE TABLE model_dataset_weight AS
            SELECT MD.*, GS.err_std, 1/(GS.err_std**2) AS weight
            FROM model_dataset AS MD INNER JOIN group_stderr AS GS
            ON MD.Group = GS.Group;
    QUIT;
```

After estimation of error variances and weights we apply SAS *proc GLM* with WEIGHT statement.

```
    PROC GLM DATA = model_dataset_weight OUTSTAT = output_anova2;
        WEIGHT weight;
        CLASS Group;
        MODEL logY = Group X1 X2 / NOINT SOLUTION;
        OUTPUT OUT = out2 PREDICTED = y_forecast2;
        ODS OUTPUT PARAMETERESTIMATES = output_estimate2;
    RUN; QUIT;
```

The following table contains parameter estimates from WLS model.



Dependent	Parameter	Estimate	StdErr	tValue	Probt
logY	Group 1	0.832193613	0.3474589	2.4	0.0173
logY	Group 2	1.079015956	0.35432445	3.05	0.0025
logY	Group 3	1.963988906	0.10261083	19.14	<.0001
logY	Group 4	2.054904585	0.28065086	7.32	<.0001
logY	Group 5	2.948159909	0.31157333	9.46	<.0001
logY	Group 6	2.693603347	0.13521914	19.92	<.0001
logY	Group 7	2.442019463	0.24026788	10.16	<.0001
logY	Group 8	2.872297824	0.26862718	10.69	<.0001
logY	Group 9	3.277382007	0.1561331	20.99	<.0001
logY	Group 10	3.280424715	0.21850829	15.01	<.0001
logY	X1	1.121038048	0.10531579	10.64	<.0001
logY	X2	-2.54853549	0.22703253	-11.23	<.0001

Table 2: Results from Weighted GLM Model

After correcting for heteroscedasticity we again predict Y in a naïve way by simply taking exponential of *y_forecast2* from the WLS results.

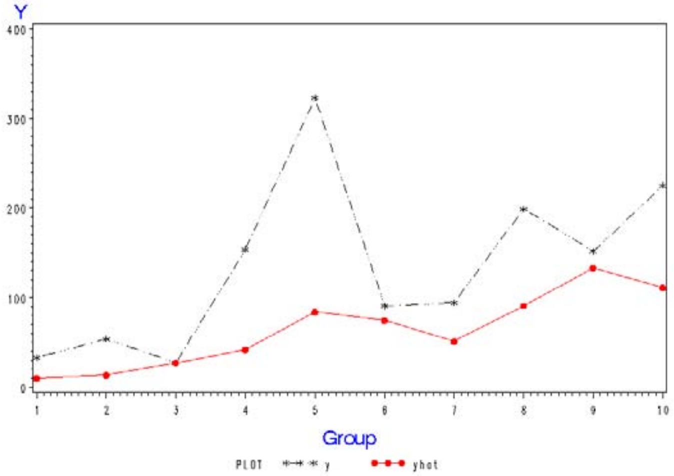


Figure 4: Naïve Anti-Log from Weighted Model

Similar to Figure 2, Figure 4 demonstrates that the naïve way once again leads to huge under-estimation. On the other hand, the plot of group means in Figure 5 illustrates the results obtained by applying adjustments discussed in Section 1 on WLS estimates in order to account for both heteroscedasticity and bias associated with anti-log transformation.

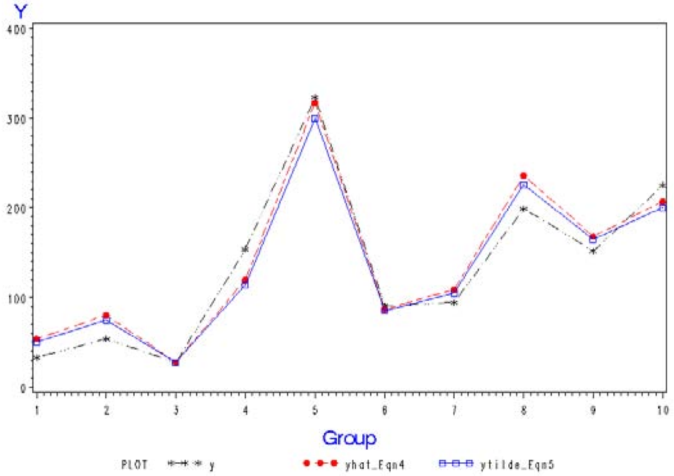


Figure 5: $\hat{Y}$ and $\tilde{Y}$ from WLS

It is apparent from Figure 5 that prediction results have improved significantly as compared to earlier results. The improvement is due to the adjustment not only accounting for bias associated with anti-log transformation but also with better variance estimates. Moreover, the prediction results from equation (5), $\tilde{Y}$, is superior to $\hat{Y}$ from equation (4). This is attributed to the fact that equation (5) corrects for bias associated with parameter estimates whereas equation (4) ignores them.

CONCLUSION

This study warrants caution while dealing with log-transformed linear models plagued with the problem of heteroscedasticity. In order to predict correctly analysts not only need to remove bias associated with anti-log transformation by applying proper adjustments as suggested in this paper but also need to correctly estimate the variance terms. The SAS system facilitated building of a routine which can easily incorporate all the adjustments mentioned in the paper.

REFERENCES

Miller, D. (1984) "Reducing Transformation Bias in Curve Fitting," *The American Statistician*, 38(2), 124–126.

Teeken, R. and Koerts, J. (pre-1986) "Some Statistical Implications of the Log Transformation of Multiplicative Models," *Econometrica*, 1972, 40(5), 793.

SAS Institute (2007) "SAS OnlineDoc 9.1.3,"
http://support.sas.com/onlinedoc/913/docMainpage.jsp.

ACKNOWLEDGMENTS

The authors thank Momentum Market Intelligence for making this study possible. The authors are also grateful to Ken Wisniewski for many helpful discussions and Alynda Lingwood for assistance in preparing graphics by SAS.

CONTACT INFORMATION

Dr. Siwei Jia
Momentum Market Intelligence
220 NW 2^{nd} Avenue, Suite 600
Portland, Oregon 97209
Tel: 503-445-4865
Email: siwei_jia@mointel.com