Paper 131-2007

Skewness, Multicollinearity, Heteroskedasticity - You Name It, Cost Data Have It! Solutions to Violations of Assumptions of Ordinary Least Squares Regression Models Using SAS®

Leonor Ayyangar
Health Economics Resource Center (HERC)
VA Palo Alto Health Care System
Menlo Park, CA

ABSTRACT

What options are available to the researcher when one or more assumptions of an ordinary least squares (OLS) regression model are violated? This poster highlights SAS® procedures that may be employed when data at hand does not fulfill the assumptions underlying OLS.

INTRODUCTION

This paper briefly describes the assumptions of the OLS regression model. SAS/STAT® Version 9.1 procedures that can be employed to test these assumptions are described and illustrated by sample codes. The consequences of violating these assumptions are enumerated. Finally, solutions are recommended.

HEALTH CARE COST DATA

Health care cost data are often messy. They have a tendency to be skewed because of very expensive events (usually a few cases that require costly procedures). It is not an uncommon practice in various other fields of research to consider these outliers as data entry errors and to exclude them from the regression model. In health economics research however, expensive events are critically important. In fact, they are considered invaluable because of their ability to shed light on what drives up health care costs.

Skewness is just one of the problems that affect the analysis of cost data. Other problems that researchers encounter include heteroskedasticity, multicollinearity and autoregressive error terms.

For example, consider cost data from an ongoing study of patients who all underwent coronary artery bypass grafting (CABG) surgery. The study aims to identify significant predictors of surgical cost. The dependent variable is total surgical cost (tsur_tot) and the independent variables are total minutes in surgery (totmin), total days in ICU (toticu), the patient's age (age), the number of pre-surgery chronic conditions (numcomplic), and blood-related variables used during surgery: the amount of packed red blood cells (RBC) and salvaged blood (savebld).

AN OVERVIEW OF THE ORDINARY LEAST SQUARES MODEL

Consider the following OLS linear model with k independent variables and n observations:

$$Y = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \ldots + B_k X_{ik} + \varepsilon_i$$

Where ε_i is the disturbance term and a stochastic component of the model.

For a given **X**, the model predicts the expected value of Y:

$$E(Y_i \mid X_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \ldots + \beta_k X_{ik}$$

Estimates of $\beta_1, \beta_2, \beta_3 \ldots\ \beta_k$ are derived using the method of least squares, which minimizes the sum of the squared deviations of the residuals**.**

PROPERTIES OF OLS ESTIMATORS

OLS estimators are BLUEs – Best, Linear, and Unbiased Estimators. They are "best" because they have the least variance among all other linear estimators. OLS estimators are also a linear function of the observations and are unbiased. That is, $E[b_i] = \beta_i$, $i = 1,2...k$.

ASSUMPTION 1: LINEARITY

The OLS regression model is linear in the parameters, meaning that we assume that the relationship between the dependent variable and the independent variables is linear.

DETECTING NONLINEARITY

Graphical techniques are useful in detecting nonlinearity. Fit the model using PROC REG and output the residuals. Note that in the case of a simple linear regression model (with one independent variable) all you need to do is create a scatter plot (using PROC PLOT) of the dependent variables and the residuals. You can also run PROC CORR to calculate correlations, in order to measure the linear association between Y and X. For multiple linear regression models, partial residual plots can detect nonlinearity. These plots are available for output in PROC REG and show the relationship between the dependent variable Y and each of the k independent variables.

Sample codes using the previously described CABG data are provided below.

Sample Code 1 Generate Partial Residual Plots

```
PROC REG data=cabgdata;
 MODEL dsur_tot = totmin  iopktrf  rbc savebld toticu age numcomplic / partial;
 TITLE' Graphical Test of Linearity Assumption';
 QUIT;
```

CONSEQUENCES OF NONLINEARITY

When the assumption of linearity is violated, OLS models will yield biased parameter estimates. In general the OLS estimators as well as R-square will be underestimated.

SOLUTIONS TO NONLINEARITY

Specification of independent variable(s)

Age is often entered as a predictor of health care costs in OLS regression models. When it is entered as a continuous variable in the model, it imposes a strong linear assumption between age and costs. This assumption may not be correct. By categorizing age into a few groups and then using dummy variables, we allow for more flexible specifications.

Transform the independent variable(s)

This is a classic textbook recommendation; however it can be problematic in real life. We can transform the independent variables using log, inverse or polynomial transformations but the interpretation of the results may be difficult to communicate. Often it is desirable to undo the transformation after regression. In this case, avoid retransformation bias by using Duan's smearing operator (Duan). If you cannot find a transformation that fits the data, spline transformations (using PROC TRANSREG) may be a good starting point.

Use a nonlinear regression model

Remember that the goal of OLS regression is to fit a straight line to the data at hand. If the functional form of the relationship between the dependent and independent variables is not linear, it is not advisable to use a linear modeling algorithm to analyze the data – the results will be unreliable. SAS/STAT® provides other modeling techniques that can be useful when an OLS model is inappropriate, most notably PROC GENMOD.

ASSUMPTION 2: INDEPENDENCE OF ERROR TERMS

OLS regression models assume that error terms are uncorrelated. This implies that observations are independent. When they are not, then there is *autocorrelation*.

DETECTING AUTOCORRELATION

Health care cost data often suffers from autocorrelation because costs tend to be calculated over time, e.g. costs that are calculated for each successive patient visit to the hospital. Spatial autocorrelation can also occur. For example, costs are assigned by each hospital and those in the same geographic location generally charge similar rates for their services.

PROC REG tests for first-order autocorrelations using the Durbin-Watson coefficient (DW). The null hypothesis is no autocorrelation. A DW value between 1.5 and 2.5 confirms the absence of first-order autocorrelation. In the examples below, timepd is the time-dependent variable. Plot the residuals against the timepd. If the error terms are independent, this graph will show only a random scattering of the residuals with no apparent pattern. The Lagrange Multiplier test is a general test of serial correlation and can be done in SAS® by re-running the regression with the lagged residuals as extra independent variables.

Sample Code 2. Test autocorrelation with the Durbin-Watson test

```
 PROC REG data=cabgdata;
  MODEL dsur_tot = totmin  rbc savebld toticu age numcomplic/dw;
  OUTPUT OUT=autocorr_test (keep= timepd res)
  RESIDUAL=res;
  TITLE’ Durbin-Watson Test of Autocorrelation';
QUIT;
```

Sample Code 3. Graphical test of autocorrelation

```
  PROC PLOT data=autocorr_test;
  PLOT RES*timepd;
  TITLE’Graphical test of autocorrelation’;
QUIT;
```

Sample Code 4. Lagrange Multiplier general test of Serial Correlation

```
 PROC REG data=in.sampledata;
  MODEL dsur_tot = totmin  rbc savebld toticu age numcomplic;
  TITLE’Output Residuals and Calculate  their Lagged Values’;
  OUTPUT OUT=outres
  RESIDUAL=res;
DATA LMTest;
  SET outres;
lagresid=lag(res);
label lagresid='lagged residual';
 PROC REG;
  MODEL res = totmin  rbc savebld toticu age numcomplic lagresid;
  TITLE’Lagrange Multiplier Test of Serial Correlation (Ho: No serial correlation)’;
QUIT;
```

CONSEQUENCES OF AUTOCORRELATION

Autocorrelation inflates t-statistics by underestimating the standard errors of the coefficients. Hypothesis testing will therefore lead to incorrect conclusions.

SOLUTIONS TO AUTOCORRELATION

Perform a thorough exploratory analysis of your data

If you detect autocorrelation, do not abandon the OLS model right away. It is not advisable to rely on just one test since that test could send mixed signals. For example, a significant DW test has been known to result from a violation of the linearity assumption or from model misspecification.

(Kennedy, p.142) Multiple tests for autocorrelation, including comparison of the results, are recommended.

Transform the dependent variable and/or the independent variable(s)
Adding lagged transforms of the dependent and/or the independent variables may get rid of autocorrelation~~.~~

Use PROC AUTOREG
If you are sure you have a legitimate case of autocorrelation, then it is advisable to use PROC AUTOREG (available in SAS/ETS®), which was developed specifically for time series data.

ASSUMPTION 3: $\varepsilon_i \sim N(0,\sigma^2)$

The errors are assumed to be normally distributed with mean zero and a homogenous variance.

DETECTING NON-NORMALITY OF ERROR TERMS

Non-normality of error terms is easily detected graphically. Run the regression using PROC REG and output the residuals. Use PROC UNIVARIATE to produce normal probability plots and PROC KDE to produce kernel density plots. PROC UNIVARIATE also outputs several tests for normality such as the Shapiro-Wilk test and the Kolmogorov-Smirnov test. The null hypothesis is the error terms are normally distributed.

Sample Code 5. Test for Normality of Error Terms

```
PROC REG data=cabgdata;
  MODEL dsur_tot = totmin  rbc savebld toticu age numcomplic;
  OUTPUT OUT=outres
  RESIDUAL=res  PREDICTED=Yhat;
QUIT;


PROC UNIVARIATE data=outres normal;
 VAR res;
 HISTOGRAM res / normal;
 PROBPLOT  res;
 TITLE'Tests for Normality of Residuals';
QUIT;
```

DETECTING HETEROSKEDASTICITY

Heteroskedasticity occurs when the error variance is not homogenous. To detect heteroskedasticity, plot the residuals against the predicted value and look for patterns in the graph. PROC REG also performs the White test for heteroskedasticity. The null hypothesis states that there is no heteroskedasticity.

Sample Code 6. Test for Heteroskedasticity

```
PROC REG data=in.cohort;
  MODEL dsur_tot = totmin  rbc savebld toticu age numcomplic/spec;
  TITLE 'White Test of Heteroskedasticity';
QUIT;
```

Keep in mind that the White test is not very discriminating – it tends to pick up only extreme cases of heteroskedasticity.

CONSEQUENCES OF NON-NORMALITY AND HETEROSKEDASTICITY

Normality of error terms is required for the statistical tests to be valid. Heteroskedasticity results in inefficient estimators and biased standard errors, rendering the t-tests and confidence intervals unreliable. As in the case of autocorrelation, hypothesis testing will lead to incorrect conclusions. The estimators however, remain unbiased.

SOLUTIONS TO NON-NORMALITY AND HETEROSKEDASTICITY

Transform the dependent variable(s)

Try using transformations to stabilize the variance. Popular transformations of the dependent variable include the square root, log and reciprocal transformations. The log transform can be problematic if you want to exponentiate the results to present the results in dollars. In this case, it is advisable to use Duan's smearing operator (Duan, 1983).

Use Weighted Least Squares

This is accessible in PROC REG by using the WEIGHT statement and specifying the weighting variable.

Use PROC ROBUSTREG

This new procedure from SAS® was developed for data characterized by outliers – PROC ROBUSTREG is a welcome alternative to PROC REG, since OLS estimates are sensitive to outliers. It attempts to "rein in" these outliers and calculate stable and resistant estimators using robust regression techniques.

ASSUMPTION 4 : MEAN INDEPENDENCE : $E[\varepsilon_i | X_{ij}]=0$

OLS assumes fixed, not random, predictors that are distributed independently of the error terms.

DETECTING VIOLATIONS OF MEAN INDEPENDENCE

Violations of mean independence are referred to as *"endogeneity"* in econometrics. This means that the independent variable is not truly independent - it was not manipulated by the experimenter. This is a very common problem with observational analyses and occurs when there are omitted variables, a recursive model and/ or measurement errors. The SPEC option in PROC REG performs a joint test of heteroskedasticity, model specification and mean independence.

CONSEQUENCES OF ENDOGENEITY

Violations of mean independence results in biased estimators and standard errors.

SOLUTIONS TO ENDOGENEITY

Use PROC SYSLIN

This procedure performs two-stage least squares regression, a method developed especially to handle endogeneity. Suppose you wanted to regress a health outcome (e.g., quality adjusted life years) on alcohol drinking. The use of alcohol is endogeneous in this model - we did not randomly assign people to drink. The first stage of PROC SYSLIN regresses the endogeneous regressor (alcohol use) on other covariates and at least one instrumental variable (state tax rate for alcohol). The instrumental variable should be a strong predictor of the endogenous variable and should not be correlated with the original dependent variable (quality adjusted life years). In the second stage, the predicted values from the first stage regression are used. We seldom use this method in practice because the existence of strong instrumental variables is very rare.

ASSUMPTION 5: X_i IS UNCORRELATED TO X_j , $i \neq j$

There is no intercorrelation among the independent variables, i.e. there is no multicollinearity.

DETECTING MULTICOLLINEARITY

SAS provides several measures of collinearity in PROC REG. You can output the variance inflation factor (VIF) or the tolerance. Look for independent variables with VIFs or condition index greater than 10 and with tolerance less than 0.10.

Sample Code 8. Test for Multicollinearity

```
PROC REG data=in.sampledata;
```

```
  MODEL tcst_tot = tsur_tot age los /vif tolerance collinoint;
  TITLE' Test for Multicollinearity';
QUIT;
```

CONSEQUENCES OF MULTICOLLINEARITY

Multicollinearity inflates the standard errors, making it impossible to determine the relative importance of the predictors. In other words, the coefficients will be unreliable. Note that multicollinearity does not affect the efficiency of the estimators – they remain BLUE.

SOLUTIONS TO MULTICOLLINEARITY

Remove all but one of the highly correlated variables from the analysis
There is no need to retain redundant variables, since they measure the same characteristic.

Use highly correlated variables as interaction effects or as a composite variable
All highly correlated variables can be deleted and re-introduced in the model in the form of an interaction or as a composite variable.

If possible, increase the sample size
Increasing sample size will decrease the standard errors, which is greatly inflated by the presence of highly intercorrelated variables.

CONCLUSION

PROC REG is a useful tool for detecting violations of the assumptions of an OLS regression model. It can output information into PROC PLOT to develop graphs that are useful in the detection of data disturbances.

In our health care cost studies, we often see violations of one or more assumptions. As such, when a decision is made to use OLS regression models, we employ a combination of the solutions described above. Note that when several violations occur, the use of OLS regression models for health care cost data is difficult to justify. However, an OLS model is a good beginning model because it is easy to understand. Furthermore, with large samples, we can employ the Central Limit Theorem to justify our choice.

It is important to know what the model's purpose is – is it descriptive or predictive? OLS regression models for CABG surgery data have performed as well as other models in identifying and examining the impact of factors associated with cost (Austin, et al). OLS models have the advantage of simplicity and clarity as well as being easy to code. These models, however, do not perform as well as other models in predicting cost for future patients.

PROC ROBUSTREG, a new procedure in SAS® was developed to handle data with outliers, and can be useful for OLS regression modeling of cost data.

REFERENCES

Austin, Peter C., et al . *A Comparison of Several Regression Models for Analysing Cost of CABG Surgery*, STATISTICS IN MEDICINE (2003); 22:2799-2815
Chen, Colin (Lin) *Robust Regression and Outlier Detection with the ROBUSTREG Procedure*, SUGI 27 Proceedings
Der, G. and B.S. Everitt (2002). *A Handbook of Statistical Analyses using SAS*, 2nd ed., Chapman & Hall/CRC.
Duan N. Smearing estimate: a nonparametric retransformation method. Journal of the American Statistical Association 1983;78:605-610.
Greene, William H. (2003). *Econometric Analysis*, 5th ed., Pearson Education, Inc.
Kennedy, Peter. (1992). *A Guide to Econometrics*. Cambridge, MA: Massachusetts Institute of Technology Press.
Neter, Wasserman, and Kunter (1990). *Applied Linear Statistical Models*, 3rd ed., Irwin.
SAS Institute. 2004. *SAS/STAT 9.1 User's Guide, Version 8*. Cary, NC: SAS Institute

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Leonor Ayyangar
Health Economics Resource Center (HERC)
Palo Alto VA Health Care System
795 Willow Road, (152 MPD)

Menlo Park, CA 94025
Phone: (650) 493-5000 Ext. 22338
E-mail: Leonor.Ayyangar@va.gov

Acknowledgement
The author would like to thank Todd Wagner for his encouragement and suggestions.
Special thanks to Vilija Joyce, Patsi Sinott and Bei Chang for taking time off from
their busy schedules to review this paper.