

Take-home Graded Practice Opportunity 6

Due date: December 6, 2011, 5:00 p.m.

Conceptual practice with Gauss-Markov assumption violations

1. Explain how serial autocorrelation may occur due to an omitted variable. Provide an example serial autocorrelation due to variable omission.

Recall that serial autocorrelation is when the error terms are related across times. If a potential explanatory variable is correlated over time (e.g., weather) and is not explicitly controlled for in the model, then it is included in the error term. Therefore, the error terms become serially correlated.

2. In non-technical terms (i.e., in a manner that individuals without an econometrics education) explain the concept of stationarity. Why is it useful? Why do we need it?

Stationarity is a concept that implies that a time-series variable can be explained forward and backward in time, because the mean and variance of the variable do not change in time. If variables are stationary, then we can describe their long-term behavior in time.

3. In the following three scenarios an OLS regression is estimated. In each scenario, are any Gauss-Markov assumptions violated? If so, which assumptions are violated and why?

- (a) A model of stock prices as a function of changing market conditions.

This is a case of heteroskedasticity and potential serial autocorrelation. Heteroskedasticity may occur because specific market conditions can be systematically associated with either high or low variability. Serial autocorrelation may occur because information across time is related, and any unobservable factors that may be driving changes in market conditions may be incorporated into the error term of the stock prices.

- (b) A model of expenditures on housing as a function of income.

This is a case of heteroskedasticity. More affluent people have a lot more opportunities to choose how to spend their money on housing. Therefore, the variability of expenditures on housing will increase with the income level.

(c) A model of median income on per-capita level of public education spending. *Similarly to (b), as the level of public education spending increases, individuals acquire a higher level of education and can have a greater opportunity to choose a career. Some will choose high paying careers and some will choose lower paying careers. Therefore, the variability will be substantial. Conversely, those attending poor quality schools will not have as many career opportunities, and their income levels are likely to be more similar.*

4. Recall that the GLS estimator can be written as: $\tilde{\beta}_{GLS} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$. Show that this estimator minimizes the weighted sum of squared errors:

$$(\mathbf{y} - \mathbf{X}\tilde{\beta})'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\tilde{\beta})$$

We only need to look at the first term: $(\mathbf{y} - \mathbf{X}\tilde{\beta})$:

$$\begin{aligned} \mathbf{y} - \mathbf{X}\tilde{\beta} &= \mathbf{y} - \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} \\ &= \mathbf{y} - \mathbf{X}\mathbf{X}^{-1}(\mathbf{X}'\mathbf{V}^{-1})^{-1}(\mathbf{X}'\mathbf{V}^{-1})\mathbf{y} \\ &= \mathbf{y} - \mathbf{I}_n\mathbf{y} \\ &= \mathbf{y} - \mathbf{y} \\ &= \mathbf{0} \end{aligned}$$

Because \mathbf{V}^{-1} is a positive semi-definite matrix, the weighted sum of squared errors will be zero at $\tilde{\beta}_{GLS}$.

Applied analysis

You are examining crop insurance policies sold in the United States in 2009. Specifically, you are focused on policies sold in U.S. counties for canola. A dataset describing this information as well as other related characteristics is located on the course website (“data_hw6.csv”) or at this web address:

http://www.montana.edu/bekkerman/classes/ecns561f11/data_hw6.csv

The dataset contains the following variables:

Variable	Description
crop_yr	Year
fips	FIPS county identifier
county_name	County name
state	State name
crop_name	Crop name
ins_name	Crop insurance policy name
cov_lvl	Level of insurance coverage (percentage)
policies_sold	Number of policies sold
net_acres	Number of acres insured
liability	Total liability (maximum potential production loss if 100% loss occurred)
total_prem	Total premiums paid by farmers to be insured
indem_amt	Amount of indemnities paid
loss_ratio	Ratio of indemnities to premiums

You would like to empirically analyze factors affecting the number of crop insurance policies sold. Your goals are the following:

1. Organize and clean the data as necessary. Provide a brief overview of your steps.
2. Manipulate the data so that you have average values across the two different insurance plans.
3. Describe the model that you would seek to estimate using the dataset. Justify each variable in your model with an explanation of how the variable is economically relevant and what it will explain.

The dependent variable is policies_sold. The independent variables are selected to describe factors that could affect the number of policies sold. These would include net acres insured, the total liability (higher potential losses are likely to increase policy sales), premiums paid (lower costs increase policy sales), and loss ratios.

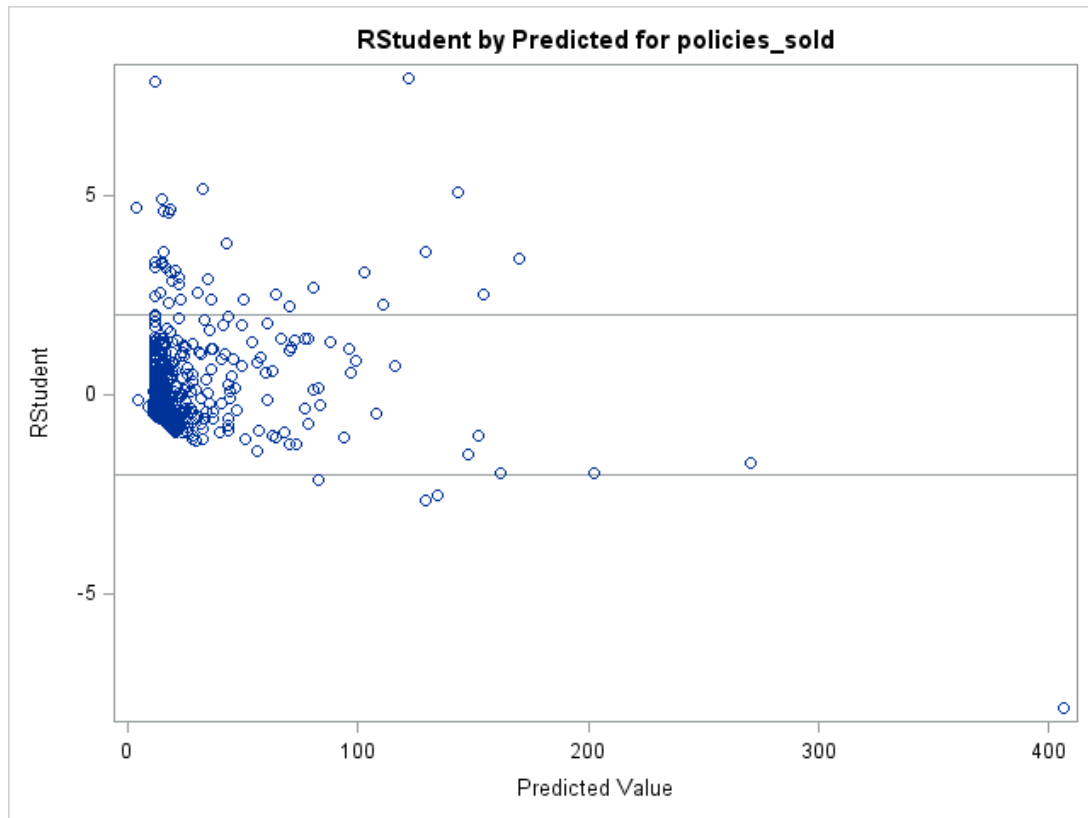
4. Provide visual and statistical summary statistics of the relevant variables. What are the potential economic implications? (*Hint: compare the mean and the median. What do they imply?*)

In comparing the mean and median, there is quite a bit of discrepancy. This implies that the mean is heavily biased upward by potential outliers. This can bias the estimation results.

5. Estimate the OLS model.

6. Empirically diagnose the data for potential problems related to outliers. Describe whether you need to deal with outliers and if so, how you dealt with them.

The studentized residuals plots indicates that there are at least two points that could potentially create skewed results.



7. Without empirically testing for it, is heteroskedasticity a potential problem in this model? If so, what might be the source?

The variance of policies sold may be closely associated with total liability and/or net acres insured. Areas with larger areas to insure are likely to have greater variability in the number of sold policies.

8. Empirically test for heteroskedasticity.

White's test of homoskedasticity indicates that the null hypothesis is rejected and suggests that heteroskedasticity may be a concern.

9. In class (and in the course notes), we discussed the procedure for estimating feasible GLS estimators. Use this procedure to estimate the heteroskedasticity-adjusted estimator.

SAS coding hints:

- (a) In PROC REG, you can output residuals using the code line: `output out=resid_data r=e_hat;` (to be placed after the `model ...;` line).
- (b) In PROC REG, you can output predicted values using the code line: `output out=pred_data p=y_hat;` (to be placed after the `model ...;` line).
- (c) In PROC REG, you can assign weights as follows:

```
proc reg data=myData;  
weight weightVariable;  
model y = x1 x2 ...;  
run;
```

10. Estimate the OLS estimator, but adjust the standard errors using White's heteroskedasticity-robust method. How do these results differ from the FGLS estimator?