Intro to Econometric Theory
Heinz School, Carnegie Mellon University
90-906, Spring 2004-5

Solution #3

1. **Consider the generalized regression model. Suppose we try to run GLS using a matrix Ω_1 which we believe is equal to (or a good estimator of) Ω. Suppose further that we are wrong, and that $\Omega_1 \neq \Omega$. Let's call the estimator we calculate using Ω_1, $\hat{\beta}_{WGLS}$ (W for wrong).**

   (a) **Is $\hat{\beta}_{WGLS}$ unbiased?**

$$\begin{aligned}\hat{\beta}_{WGLS} &= (X'\Omega_1^{-1}X)^{-1}X'\Omega_1^{-1}Y \\ &= (X'\Omega_1^{-1}X)^{-1}X'\Omega_1^{-1}(X\beta + \epsilon) \\ &= (X'\Omega_1^{-1}X)^{-1}X'\Omega_1^{-1}X\beta \\ &+ (X'\Omega_1^{-1}X)^{-1}X'\Omega_1^{-1}\epsilon) \\ &= \beta + (X'\Omega_1^{-1}X)^{-1}X'\Omega_1^{-1}\epsilon)\end{aligned}$$

   Clearly, $\hat{\beta}_{WGLS}$ is unbiased, since we can just pass an expectation conditional on X all the way in to hit ϵ to make the second term zero.
   Wrongly correcting heteroskedasticity with GLS does not render the wrong GLS estimator biased.

   (b) **What is its variance?**

$$\begin{aligned}V(\hat{\beta}_{WGLS}) &= V((X'\Omega_1^{-1}X)^{-1}X'\Omega_1^{-1}\epsilon) \\ &= (X'\Omega_1^{-1}X)^{-1}X'\Omega_1^{-1}V(\epsilon)\Omega_1^{-1}X(X'\Omega_1^{-1}X)^{-1} \\ &= \sigma^2(X'\Omega_1^{-1}X)^{-1}X'\Omega_1^{-1}\Omega\Omega_1^{-1}X(X'\Omega_1^{-1}X)^{-1}\end{aligned}$$

   (c) **Can you think of a reasonable estimator for this variance?**

$$\hat{V}(\hat{\beta}_{WGLS}) = (X'\Omega_1^{-1}X)^{-1}X'\Omega_1^{-1}[\text{diag}(e'e)]\Omega_1^{-1}X(X'\Omega_1^{-1}X)^{-1}$$

(d) **Is this variance bigger or smaller than the GLS variance?**
Bigger, GLS is efficient.

(e) **What should we conclude about the dangers of trying to correct for a non-scalar covariance matrix, but screwing it up?**
Probably not too dangerous. Will not introduce bias/inconsistency. Will not achieve GLS efficiency, but neither will OLS. If we are worried about not having got it right, we can always use Huber/White std errors on the "wrong" GLS model. Asymptotically, it seems there is only upside to trying to fix heteroskedasticity. In small samples, of course, we could be making things worse.

2. **With respect to the hospital cost dataset from the web site, please estimate the following two models:**

$$
\begin{aligned}
lnC &= \beta_1 + \beta_2 lnD + \beta_3 lnV + \beta_4 time + \epsilon && (1)\\
C &= \alpha_1 + \alpha_2 D + \alpha_3 V + \alpha_4 time + \epsilon && (2)
\end{aligned}
$$

In the above, D is inpatient days, V is outpatient visits, and $time$ is the number of quarters since the first quarter of 1991.

(a) **Discuss how the estimates that these two equations provide for marginal costs differ.**
See the SAS program and the output. The marginal cost in the linear model may be simply read off the output. A visit costs \$1,490 and a day costs \$202.
The marginal cost calculation for the log - log model is more involved. Marginal cost is different for each observation, due to non-linearity, so there is no single "marginal cost" as in the linear model. The best we can do to deliver a single number is probably to take averages across the data. That is done in the SAS program and output. On page 5, we see that the estimates for marginal cost are \$768 per visit and \$360 per day, respectively.

(b) **Test for non-linearity in equation 2. Discuss the meaning of your findings.**
I used RESET test, see code. To do the test, compare regressions on pages 6 and 1. $SSE_{UR} = 317416$ $SSE_R = 379003$.

$$
\begin{aligned}
\text{F-stat} &= \frac{(SSE_R - SSE_{UR})/2}{SSE_{UR}/(N-K)}\\
&= \frac{(379-317)/2}{371/7536}\\
&= 1,220,212
\end{aligned}
$$

We will reject the null of no non-linearity with great confidence. You can also see this with the huge t-stats on the Ramsey terms in the regression on page 6.

(c) **Think carefully about the variable C. Think carefully about equation 2 (not the estimated version, but the theoretical version). Are there circumstances in which equation 2 can give obviously wrong predictions for$C$? Does equation 1 suffer from a similar problem?**

Cost must be always positive. The linear specification can give negative predicted values. The log specification cannot give negative predicted values for cost.

(d) **Calculate predicted values for C from both models. Calculate the mean squared error for both models: $\sum(C - \hat{C})$. Which model is better by this criterion?**

MSE for the linear model is on page 1. It is 50.3.

The calculation of MSE for the log-log model is involved and requires the use of the smearing estimator. It is detailed in the SAS program. The MSE of the log-log model is 59.5. So, the log-log model looks worse. (Notice, the G-M Thm does NOT say that this has to happen, since the log-log model is not linear!)

(e) **Would it make sense to compare the R^2 from the two equations in order to decide which is best? Why or why not?**

No. The two regressions have different LHS variables. R^2 is only meaningful for comparisons of regressions with the same LHS variables.

(f) **If we care about getting the right answer, which equation should we use and why? (You may do more analysis if you like to answer this question)**

Well, the RESET test rejected the linear specification, and theoretically the linear specification is undesirable since it could yield negative predicted values for costs. Looking on page 9 of the output, we see that the linear specification indeed does give some negative predicted values for costs, which is pretty disturbing. On the other hand, the log-log model gives worse MSE in the level of costs.

On balance, I would probably use the log-log model unless I really do have a loss function which is given by the mean square prediction error for costs.

In reality, I would probably work on the log-log model to improve it. We could test it for non-linearity and then include, say, squared and interaction terms. (the log-log cost function with squared and interaction terms is called a trans-log cost function).

3. The Medical Expenditure Panel Survey is an annual survey which collects information about medical expenditures, income, employment, demographics, health information, &c for a representative sample of Americans.

   I have prepared an extract of these data for 1996, and it is available on the course website. The following are the columns in the data, in order:

Variable	Meaning
age	age of person in years
sex	sex of person, 1=male & 0=female
income	income in 1996 $
employed	1=employed, 0=not employed
insured	1=had health insurance, 0=not
health	perceived health status, higher is sicker
spending	spending on health care, 1996 $

   To begin with, let's consider this model:

$$\begin{aligned} \text{spending}_i &= \beta_1 + \beta_2 \text{income} + \beta_3 \text{age} + \beta_4 \text{sex} \\ &+ \beta_4 \text{employed} + \beta_5 \text{insured} + \beta_6 \text{health} \end{aligned} \tag{3}$$

   (a) **Test this model for heteroskedasticity. Discuss your findings and and what they mean for inference.**

   See page 11 of the output. This shows the results of White's test, which rejects homoskedasticity at $p < 0.0001$. So, although the OLS estimates are unbiased, the displayed standard errors are wrong.

   (b) **How much more do people in very good health status spend than do people in excellent health status (estimate and CI). Make sure your answer is correct in the presence of heteroskedasticity.**

   By careful choice of which dummies to include, we can just read the estimate off the output on page 10. It is $286. To make the CI, we need to use heteroskedasticity robust standard errors. The heteroskedasticity robust variance matrix appears beginning on page 11. The variance of β_{hsvryg} is 8678. Thus, the 95% CI is $286 \pm 1.96\sqrt{8678}$, or 286 ± 183.

   (c) **Test whether insurance affects spending for people of different health statuses differently and discuss. Make sure your answer is correct in the presence of heteroskedasticity.**

Answer appears on page 18. We reject he null that health status does not interact with insurance status.

4. **Do problems 5 and 6 on pp 248-9 of Greene.**

   Answer appears on pages 20-32 of the output. White's test does not reject homoskedasticity (page 24). The Breuch Pagan regression appears on page 25, and its statistic is $\frac{1}{2}128 = 64$, and this obliterates the table values for chi_2^2. The regressions for the Goldfeld-Quandt test using $X1$ are on pages 26 and 27, and the F-statistic is $\frac{5.62}{3.91} = 1.44$. The regressions for the Goldfeld-Quandt test using $X2$ appear on pages 28 and 29 and the F-stat is $\frac{8.42}{4.08} = 2.06$. Both tests should be referred to the F table with 20 num and 20 denom degrees of freedom. The table value is 2.12, so both miss significance at the 5% level.

   The weighted least squares estimation appears on page 32.