EXERCISE ON GLS, VARIABLE TRANSFORMATION

We are going to look at whether using GLS on the `caschool` test score regression is useful, and at whether it might affect our conclusions.

Start by estimating a linear regression of the `caschool` variable `testscr` on `str`, `comp_stu`, `meal_pct`, `calw_pct`, `avginc`, `el_pct` and `teachers`.

(1) Compute the White (heteroskedasticity-robust) standard errors for the coefficients and compare them to those implied by the SNLM. In R, you can get the White standard errors from the diagonal of the covariance matrix calculated by the `hccm` command in the `car` package. This provides the covariance matrix; the square root of its diagonal is the vector of estimated coefficient standard errors.

(2) Estimate a regression of the squared residuals from the regression in (1) on `1/teachers`. This seems like a reasonable idea, because `teachers` varies from less than 10 to over 1000, and we might expect the variance of average test scores as measures of true school means to be proportional to one over the number of students. Also, if teachers vary in quality, schools with small numbers of teachers might have higher test score variability, apart from the number of students being averaged over. Prepare a scatter plot with `teachers` on the horizontal axis and squared residuals on the vertical axis. On the same plot, show the fitted values from this regression of squared residuals on `1/teachers` plotted against `teachers`. [In R, the scatter plot is just a standard `plot()`, after which you can use `lines()` to draw the fitted values (`lmresout$fitted` if your initial regression output is in `lmresout`). The lines command will not produce nice results unless you sort its inputs. Instead you can just use `points()` with the options `pch="x", col="red"` or something similar, so the fitted values are easy to distinguish. To do the sorting you would need to do `o <- order(teachers)` and `lines(teachers[o], lmresout$fitted[o])`.

(3) Estimate the same equation as in (1, but as a weighted regression, weighting squared errors by the inverse of the fitted values from your regression in (2). Explain why this approximately implements generalized least squares (GLS). [In R, reapply the `lm()` command from the (1) regression, adding the argument `weights=1/lmresout$fitted`.]

(4) Though there is no particular reason here to think this is a good idea, as part of this exercise proceed to estimate a regression of the natural log of `testscr` on the same right-hand-side variables as in the regressions in (1).

(5) Compare the coefficients and standard errors that have emerged from these four approaches: SNLM theory for the OLS estimates, OLS estimates with White standard errors, feasible GLS, and the regression with logged dependent variable. Might a researcher's conclusions about the importance of `str` depend on which approach she took?

(6) In deciding which results to base decisions on, we need to compare the fit of the regressions and the evidence for heteroskedasticity. Does the regression with squared residuals as dependent variable show strong evidence of heteroskedasticity? Does this suggest preferring White standard errors to the SNLM standard errors for the OLS estimates? The fits of the OLS, GLS and logged regressions cannot be compared by looking at R^2 or at residual standard errors. The output of R's `summary(lmout)` command gives R^2 and residual standard errors for the *weighted* data, and as we discussed in class, the units of measurement for residuals are different, and vary with the level of `testscr`, when we use logged data.

However, we can compare fit. The log likelihood (which is also of course a flat-prior posterior pdf), with σ^2 integrated out, for a SNLM estimation, is a constant (that does not vary so long as the

sample size and the explained variable does not change) plus

$$(*) \qquad -\frac{n}{2}\log\left(\frac{\sum \hat{\varepsilon}_i^2}{2}\right) .$$

When we take a non-linear transformation of the data, we have to take account of Jacobian terms, as discussed in class, so for the model with logged test score on the left-hand side, we add to the formula above, computed with the residuals from the logged dependent variable regression,

$$-\sum \log(y_i) ,$$

where in our case y_i is `testscr`. The residuals used in the first part of the formula are the differences between $\log(y_i)$ and its predicted value from the regression.

When we use weighted least squares, we again need to correct for the scale change and reweighting of data. If we start with $(*)$, calculated from the weighted residuals, then we need to add to it

$$\sum \log(w_i) ,$$

where the w_i terms are the weights. In our application, this means we *subtract* the sum of the logs of the fitted values from the squared-error regression (since the weights were the inverse of the fitted values).

Using these likelihood based measures, compare the fits of the OLS, GLS, and logged-dependent-variable models. In this application, does the use of GLS, rather than White-standard-errors and OLS, make a possible difference in conclusions about `str`'s effect?

[Note: The log likelihoods probably come out negative. This is expected. Only the differences in likelihoods matter, so a likelihood that is small in absolute value but negative is better than one that is larger in absolute value and negative.]