Solutions to Problem Set 2 – KAN-CCMVV2401U

1 Parts (b) and (c). The homoskedasticity assumption playes no role in showing that OLS is consistent. But we know that heteroskedasticity causes statistical inference based on the usual *t* and *F* statistics to be invalid, even in large samples. As heteroskedasticity is a violation of the Gauss-Markov assumptions, OLS is no longer BLUE.

2 (a) Robust standard errors are slightly larger than the usual standard errors but no change in the level of significance (except for the constant).

(b) After estimating the equation, we obtain the squared OLS residuals $\hat{u}^2$. The full-blown White test is based on the *R*-squared from the auxiliary regression (with an intercept),

$$\hat{u}^2 \text{ on } llotsize,\ lsqrft,\ bdrms,\ llotsize^2,\ lsqrft^2,\ bdrms^2,$$
$$llotsize \cdot lsqrft,\ llotsize \cdot bdrms, \text{ and } lsqrft \cdot bdrms,$$

where "*l* " in front of *lotsize* and *sqrft* denotes the natural log. With 88 observations the *F* version of the White statistic is 1.05, and this is the outcome of an (approximately) F_(9,78) random variable. The *p*-value is about .41, which provides little evidence against the homoskedasticity assumption.

3 **(a)** See output.

(b) There is functional form misspecification if $\beta_6 \neq 0$ or $\beta_7 \neq 0$, where these are the population parameters on $ceoten^2$ and $comten^2$, respectively. Therefore, we test the joint significance of these variables using the *R*-squared form of the *F* test: $F = [(.375 - .353)/(1 - .375)][(177 - 8)/2] \approx 2.97$. The *p*-value is slightly above .05, which is reasonable evidence of functional form misspecification. (Of course, whether this has a practical impact on the estimated partial effects for various levels of the explanatory variables is a different matter.)

(c) Alternatively, we can apply a RESET test to the restricted model.The p-value is ~0.01. Thus, there is some evidence for misspecification but it is still to be explored due to which variable.

4 (a) The estimated equation is

$$\underset{}{sat} = \underset{(6.20)}{997.98} + \underset{(3.99)}{19.81\ hsize} - \underset{(0.55)}{2.13\ hsize^2}$$

$$n = 4{,}137, \quad R^2 = .0076.$$

The quadratic term is very statistically significant, with *t* statistic ≈ -3.88.

(b) We want the value of *hsize*, say *hsize**, where $\hat{sat}$ reaches its maximum. This is the turning point in the parabola, which we calculate as *hsize** = 19.81/[2(2.13)] ≈ 4.65. Since *hsize* is in 100s, this means 465 students is the "optimal" class size. Of course, the very small *R*-squared shows that class size explains only a tiny amount of the variation in SAT score.

(c) With log(*sat*) as the dependent variable we get

$$\begin{array}{lll} \log(sat) = 6.896 & +.0196\, hsize & -.00209\, hsize^2 \\ \quad\quad (0.006) & (.0040) & (.00054) \end{array}$$

$n = 4{,}137, \quad R^2 = .0078.$

The optimal class size is now estimated as about 469, which is very close to what we obtained with the level-level model.

(d) The coefficients on the regressors are smaller in part (c): by factor 100 for size and by factor 10000 for sizesq. R-squared, F- and t-values are unchanged as expected. The optimal size is now *size** = 0.1981/[2(0.000213)] ≈ 465 students, which is the same as in part (b).