Solutions to Problem Set 5 – KAN-CCMVV2401U

1 (a) A possible model is:

$$y_{it} = \beta_0 female + \sum_{s=0}^{10} \alpha_s dur_{sit} + \sum_{l=1}^{4} \sum_{s=0}^{10} \gamma_{ls} dur_{sit} * \log o_{lit} + v_{it}$$

with y_{it} is one of the three physical responses, i=1,…,27 and t=1,…,110. α_s captures the base effect of the distractors at duration *s*. γ_{ls} captures the effect of logo *l* at duration *s*. There is no constant as otherwise there would be multicollinearity with base effect of the distractors.

(b) The POLS results for the model in a) suggest that most of the alphas and gammas are insignificant. While the alphas are small, there is no indication that the distractors cause some sort of physical expression. For the logos, some of the cofficients partly economically sizable but in most cases there is a lack of statistical significance. Robust standard errors are obtained to account for possible serial correlation and heteroscedasticity. A reason for the lack of statistical significance is that there are only 27 individuals (despite that there are more than 2000 observations in the panel data). Despite that a couple of individuals respond to each of the logos (compare figures), the amount of information is not sufficient to infer that the stimuli have an effect.

(c) An unobserved effects model is estimated to allow for possible correlation between the variables of interest and the time constant part of the error. Given that the key variables are exogenously set, they should not be correlated with any of the unobserved factors. This is confirmed by the FD regression results because they are similar to the POLS results. FD should be preferred over FE in this example because the number of individuals is small. The FE estimator requires N large for having desirable properties (small sample properties are sometimes not good), while the distribution of the FD estimator has desirable properties for N small and T large (as in this example). Stratifying by gender partly changes coefficient patterns but not lack of significance. Stratification may produce interesting insights for larger N. The analysis would benefit from additional data from more individuals.

2 (a) Pooling across semesters and using OLS gives

$$\begin{aligned} \widehat{trmgpa} = &-1.75 \underset{(0.35)}{} - \underset{(.048)}{.058\, spring} + \underset{(.00015)}{.00170\, sat} - \underset{(.0010)}{.0087\, hsperc} \\ &+ \underset{(.052)}{.350\, female} - \underset{(.123)}{.254\, black} - \underset{(.117)}{.023\, white} - \underset{(.076)}{.035\, frstsem} \\ &- \underset{(.00073)}{.00034\, tothrs} + \underset{(0.104)}{1.048\, crsgpa} - \underset{(.049)}{.027\, season} \end{aligned}$$

$n = 732, \quad R^2 = .478, \quad \bar{R}^2 = .470.$

The coefficient on *season* implies that, other things fixed, an athlete's term GPA is about .027 points lower when his/her sport is in season. On a four point scale, this a modest effect (although it accumulates over four years of athletic eligibility). However, the estimate is not statistically significant (*t* statistic $\approx -.55$).

(b) The quick answer is that if omitted ability is correlated with *season* then, as we know, OLS is biased and inconsistent. The fact that we are pooling across two semesters does not change that basic point.

If we think harder, the direction of the bias is not clear, and this is where pooling across semesters plays a role. First, suppose we used only the fall term, when football is in season. Then the error term and season would be negatively correlated, which produces a downward bias in the OLS estimator of β_{season}. Because β_{season} is hypothesized to be negative, an OLS regression using only the fall data produces a downward biased estimator. [When just the fall data are used, $\hat{\beta}_{season} = -.116$ (se = .084), which is in the direction of more bias.] However, if we use just the spring semester, the bias is in the opposite direction because ability and season would be positive correlated (more academically able athletes are in season in the spring). In fact, using just the spring semester gives $\hat{\beta}_{season} = .00089$ (se = .06480), which is practically and statistically equal to zero. When we pool the two semesters we cannot, with a much more detailed analysis, determine which bias will dominate.

(c) The variables *sat*, *hsperc*, *female*, *black*, and *white* all drop out because they do not vary by semester. The intercept in the first-differenced equation is the intercept for the spring. We have

$$\begin{array}{lllll} \Delta trmgpa = -.237 & +.019\,\Delta frstsem & +.012\,\Delta tothrs & +1.136\,\Delta crsgpa & -.065\,\Delta season \\ \quad\quad (.206) & \quad (.069) & \quad\quad (.014) & \quad\quad (0.119) & \quad\quad (.043) \end{array}$$

$n = 366, \quad R^2 = .208, \quad \bar{R}^2 = .199.$

Interestingly, the in-season effect is larger now: term GPA is estimated to be about .065 points lower in a semester that the sport is in-season. The *t* statistic is about –1.51, which gives a one-sided *p*-value of about .065.

(d) One possibility is a measure of course load. If some fraction of student-athletes take a lighter load during the season (for those sports that have a true season), then term GPAs may tend to be higher, other things equal. This would bias the results away from finding an effect of *season* on term GPA.

3 (a) The pooled OLS estimate of β_1 is about .360. If $\Delta concen = .10$ then $\Delta lfare = .360(.10) = .036$, which means air fare is estimated to be about 3.6% higher.

(b) The 95% CI obtained using the usual OLS standard error is .301 to .419. But the validity of this standard error requires the composite error to have no serial correlation, which effectively means a_i is not in the equation. The fully robust 95% CI, which allows any kind of serial

correlation over the four years (and any kind of heteroskedasticity), is .245 to .475 – quite a bit wider than the usual CI. The wider CI is appropriate, as the neglected serial correlation introduces uncertainty into our parameter estimators.

(c) The quadratic has a U-shape, and the turning point is about $.902/[2(.103)] \approx 4.38$. This is the value of log(*dist*) where the slope becomes positive. The value of *dist* is exp(4.38), or about 80. The shortest distance in the data set is 95 miles, so the turning point is outside the range of the data (a good thing in this case). What is being captured is an increasing elasticity of *fare* with respect to *dist* as fare increases.

(d) The FE estimate is .169, which is much lower than the pooled OLS estimate.

(e) Factors about the cities near the two airports on a route could affect demand for air travel, such as population, education levels, types of employers, and so on. Of course, each of these can be time-varying, although, over a short stretch of time, they might be roughly constant. The quality of the freeway system and access to trains, along with geographical features (is the city near a river?) would roughly be time-constant. These could certainly be correlated with concentration.