Problemløsninger PS6 – BA-BMECV1031U

1 (a) Let P(y=1 |**x**)=**xβ**, where x_1=1. Then for each *i*, ι_i(**β**)=y_i*log(**x**_i**β**)+(1-y_i)log(1-**x**_i**β**), which is only well defined for 0<**x**_i**β**<1.
(b) For any possible estimate $\tilde{\boldsymbol{\beta}}$, the log-likelihood function is well-defined only if 0<**x**_i $\tilde{\boldsymbol{\beta}}$ <1 for all i. Therefore, during the iterations to obtain the MLE, this condition must be checked. It may be impossible to find an estimate that satisfies these inequalities for every observation, especially if the number of observations is large.
(c) This follows from the Kullback-Leibler Information Criterion (KLIC): the true density of *y* given **x** – evaluated at the true values, of course – maximises the KLIC. Because the MLEs are consistent for the unknown parameters, asymptotically the true density will produce the highest average log-likelihood function. So, just as we can use an R-squared to choose among different functional forms for E(y |**x**), we can use values of the log-likelihood to choose among different models for P(y=1 |**x**) when y is binary.

2 (a) About .392, or 39.2%.

(b) The estimated equation is

= -.506+ .0124 inc - .000062 inc2+ .0265 age - .00031 age2 - .0035 male
(.081) (.0006) (.000005) (.0039) (.00005) (.0121)
n = 9,275, R2 = .094.

(c) 401(k) eligibility clearly depends on income and age in part (b). Each of the four terms involving inc and age have very significant t statistics. On the other hand, once income and age are controlled for, there seems to be no difference in eligibility by gender. The coefficient on male is very small – at given income and age, males are estimated to have a .0035 lower probability of being 401(k) eligible – and it has a very small t statistic.

(d) Somewhat surprisingly, out of 9,275 fitted values, none is outside the interval [0,1]. The smallest fitted value is about .030 and the largest is about .697. This means one theoretical problem with the LPM – the possibility of generating silly probability estimates – does not materialize in this application.

(e) Using the given rule, 2,460 families are predicted to be eligible for a 401(k) plan.

(f) Of the 5,638 families actually ineligible for a 401(k) plan, about 81.7 are correctly predicted not to be eligible. Of the 3,637 families actually eligible, only 39.3 percent are correctly predicted to be eligible.

(g) The overall percent correctly predicted is a weighted average of the two percentages obtained in part (f). As we saw there, the model does a good job of predicting when a family is ineligible. Unfortunately, it does less well – predicting correctly less than 40% of the time – in predicting that a family is eligible for a 401(k).

3 We need to compute the estimated probability first at *hsGPA* = 3.0, *SAT* = 1,200, and *study* = 10 and subtract this from the estimated probability with *hsGPA* = 3.0, *SAT* = 1,200, and *study* = 5. To obtain the first probability, we start by computing the linear function inside $\Lambda(\cdot)$: $-1.17 + .24(3.0) + .00058(1,200) + .073(10) = .976$. Next, we plug this into the logit function: $\exp(.976)/[1 + \exp(.976)] \approx .726$. This is the estimated probability that a student-athlete with the given characteristics graduates in five years.

For the student-athlete who attended study hall five hours a week, we compute $-1.17 + .24(3.0) + .00058(1,200) + .073(5) = .611$. Evaluating the logit function at this value gives $\exp(.611)/[1 + \exp(.611)] \approx .648$. Therefore, the difference in estimated probabilities is $.726 - .648 = .078$, or just under .10. [Note how far off the calculation would be if we simply use the coefficient on *study* to conclude that the difference in probabilities is $.073(10 - 5) = .365$.]

4 (a) The heteroscedastic standard errors are similar to the standard OLS standard errors. No changes in statistical significance.
The estimated probability reduces by 7.8 percentage points when pcnv increases by 0.5.

(b) avgsen and tottime are both individually insignificant. To test joint significance in the case of OLS, one can use the standard F statistic based on the R^2. In the case of heteroscedasticity robust standard errors, the usual F statistic is not valid but a rboust version of the F test is available. In both cases the p-value is ~0.84 , which suggests that the variables are jointly insignificant.

(c) The estimated probability reduces by 10 percentage points when pcnv increases by 0.5. This is slightly more than in the case of the LPM. Please note that in the probit model, the estimated change depends on the chosen value of x. For this reason, it may well be different for other values of x.

(d) PCP is 0.73. It is 0.97 if arr86=0 and 0.10 if arr86=1. This suggests that the event =1 is not well predicted. The overall percent correctly predicted is quite high, but we cannot very well predict the outcome we would most like to predict.

(e) All three variables are individually significant at the 5% level.
Both Wald and LR test suggest that the variables are jointly significant (p-value =0).
The relationship between the probability of arrest and pcnv is nonlinear because of two reasons:
i) the probit model is nonlinear by nature due to having the normal c.d.f. as link function G.
ii) pcnv enters as a quadaratic function in this model.
The sign of the estimated effect is determined by the sign of (beta_j +2*beta_h*pcnv). Therefore for small values of pcnv, the probability is increasing in pcnv, while for pcnv>0.2168/2*0.8571=0.127 it is decreasing. This means that there is an estimated deterrent effect over most of the range of pcnv.