Problem set 10

5.

- ptime86: Spending one more month in prison reduces the predicted number of arrests by 0.04.
- qemp86: Another quarter in legal employment lowers the expected number of arrests by 0.095. If 100 more individuals were employed for an additional month the predicted number of arrests will fall by 10.
- pcnv: An increase in the proportion of convictions lowers the predicted number of arrests. In particular, if we increase pcnv by 0.5, then holding all other factors fixed the predicted number of arrests reduces by -0.1303*(0.5)=-0.065. For instance, if we increase the proportion of convictions by 0.5 then we will expect to see a fall of 6.5 arrests in 100 men.
- avgsen: Increasing the average sentence length by one month increases the expected number of arrests by 0.004. However, this is a very small and insignificant effect.
- black: Being black, relative to being white, increases the expected number of arrests by 0.33.
- hispan: Being hispanic, relative to being white, increases the expected number of arrests by 0.20.

6.

- 27.7% of the individuals in the sample were arrested in 1986.
- Given our dependent variable arr is a dummy variables means this is a linear probability model (LPM). In such a model our expected (predicted) values are interpreted as probabilities.
- To see this note: $A_i = \alpha + \beta_1 ptime86_i + \beta_2 qemp86_i + \beta 3 pcnv_i + \beta_4 avgsen_i + \beta_5 black_i + \beta_6 Hispanic_i + \epsilon_i$. Therefore, $E[A_i|X] = \alpha + \beta_1 ptime86_i + \beta_2 qemp86_i + \beta 3 pcnv_i + \beta_4 avgsen_i + \beta_5 black_i + \beta_6 Hispanic_i$. Further note, given $A_i = 1$ or $A_i = 0$, we have $E[A_i|X] = 1 * P(A = 1|X) + 0 * P(A = 0|X) = P(A = 1|X)$. In words the expected value of A_i is equal to the probability A_i takes the value 1. For convience we will write this as $E[A|X] = p_i$.
- ptime86: Spending one more month in prison reduces the predicted probability of arrest by 2.4 percentage points.

- qemp86: Another quarter in legal employment lowers the expected probability of arrest by 3. percentage points.
- pcnv: Increasing pcnv by 0.5, then holding all other factors fixed the predicted probability of arrest reduces by [-0.1521*(0.5)] by 7.6 percentage points.
- avgsen: Increasing the average sentence length by one month increases the expected probability of arrest by 0.1 percentage points. However, this is a very small and insignificant effect.
- black: Being black, relative to being white, increases the predicted probability of arrest by 17 percentage points.
- hispan: Being hispanic, relative to being white, increases the expected probability of arrest by 9.6 percentage points.

Note there are three main problems with the LPM:

- **1. It is possible for it to predict probabilities outside of the 0 1 interval.**
- **2. The error terms are non-normal.**
- **3. The error terms are heteroskedastic.**
- **Point 1. is a conceptual problem, while points 2 and 3 mean the standard errors need attention, as explained below.**

7. Only 8 values lie outside the zero one interval: not too much of a concern.

8. The error term is defined as: $\epsilon_i = A_i - E[A_i|X]$, which in the LPM case (where A_i takes values 0 or 1, and the expected mean is just the probability) means $\epsilon_i = 1 - p_i$, or $\epsilon_i = -p_i$. Thus, the distribution has two central points, one at $1 - p_i$ and one at p_i. Therefore, the error term is non-normal, and in particular, it is bi-modal.

The residuals (the sample estimate of the errors) have a bimodal distribution, which is non-normal. This implies there is non-normality in the errors (as explained above), which in turn implies our t and F statistics do not follow exact t and F-distributions. However, the consequences for our model are not too severe, since we have over 2000 observations we can invoke the CLT: our t-statistics and F-statistics will follow approximate t and F distributions respectively.

9. Error terms are said to exhibit heteroskedasticity when the assumption $V(\epsilon|X) = \sigma^2$ fails. In the LPM model this must be the case. To see this note $V(\epsilon) = E[(\epsilon_i - E[\epsilon])^2] = E[(\epsilon_i)^2]$, since $E[\epsilon_i] = 0$, as we have an intercept in the model. Therefore, $V(\epsilon) = E(\epsilon^2) = (1 - p_i)^2 P(A = 1|X) - p_i^2 P(A = 0|X) = (1 - p_i)^2 p_i - p_i^2(1 - p_i) = p_i(1 - p_i)$. Given the p_i

is a function of the X's in the model, the error terms varies directly with the X's, therefore it exhibits heteroskedasticity. To test for this in the sample, run the following regression: $e_i^2 = \delta_0 + \delta_1 ptime86_i + \delta_2 qemp86_i + \delta_3 pcnv_i + \delta_4 avgsen_i + \delta_5 black_i + \delta_6 Hispanic_i + \eta_i$, and test $H_0 : \delta_1 = \delta_2 = \delta_3 = \delta_4 = \delta_5 = \delta_6 = 0$, using an F-test. The null hypothesis is homoskedasticity. As seen in the do file, we reject in favour of heteroskedasticity. The consequence is that our t-statistics and F-statistics do not follow t and F distributions respectively. The solution is to use heteroskedastic robust standard errors (formula in lecture notes). These are typically just referred to as robust standard errors, or just robust errors.

10. The coefficients are exactly the same in each model. There are marginal differences in the standard errors, typically, in the third or fourth decimal place. The differences cause no practical changes, i.e. all variables which were previously significant (insignificant) are still significant (insignificant).

11. $t = \frac{-0.0241}{0.0029}$ Note: 1. it is equal to that in the output table 2. thus, using robust standard errors hasn't changed anything about the way we calculate t-statistics. The only difference is we use the robust s.e. in place of the non-robust s.e.