7.3 (i) The t statistic on $hsize^2$ is over four in absolute value, so there is very strong evidence that it belongs in the equation. We obtain this by finding the turnaround point; this is the value of $hsize$ that maximizes $\hat{sat}$ (other things fixed): $19.3/(2\cdot 2.19)\approx 4.41$. Because $hsize$ is measured in hundreds, the optimal size of graduating class is about 441.

(ii) This is given by the coefficient on *female* (since $black = 0$): nonblack females have SAT scores about 45 points lower than nonblack males. The t statistic is about –10.51, so the difference is very statistically significant. (The very large sample size certainly contributes to the statistical significance.)

(iii) Because $female = 0$, the coefficient on *black* implies that a black male has an estimated SAT score almost 170 points less than a comparable nonblack male. The t statistic is over 13 in absolute value, so we easily reject the hypothesis that there is no ceteris paribus difference.

(iv) We plug in $black = 1$, $female = 1$ for black females and $black = 0$ and $female = 1$ for nonblack females. The difference is therefore $-169.81 + 62.31 = -107.50$. Because the estimate depends on two coefficients, we cannot construct a t statistic from the information given. The easiest approach is to define dummy variables for three of the four race/gender categories and choose nonblack females as the base group. We can then obtain the t statistic we want as the coefficient on the black female dummy variable.

7.8 (i) We want to have a constant semi-elasticity model, so a standard wage equation with marijuana usage included would be

$$\log(wage) = \beta_0 + \beta_1 usage + \beta_2 educ + \beta_3 exper + \beta_4 exper^2 + \beta_5 female + u.$$

Then $100\cdot\beta_1$ is the approximate percentage change in *wage* when marijuana usage increases by one time per month.

(ii) We would add an interaction term in *female* and *usage*:

$$\log(wage) = \beta_0 + \beta_1 usage + \beta_2 educ + \beta_3 exper + \beta_4 exper^2 + \beta_5 female + \beta_6 female\cdot usage + u.$$

The null hypothesis that the effect of marijuana usage does not differ by gender is $H_0\text{: } \beta_6 = 0$.

(iii) We take the base group to be nonuser. Then we need dummy variables for the other three groups: *lghtuser*, *moduser*, and *hvyuser*. Assuming no interactive effect with gender, the model would be

$$\log(wage) = \beta_0 + \delta_1 lghtuser + \delta_2 moduser + \delta_3 hvyuser + \beta_2 educ + \beta_3 exper + \beta_4 exper^2 + \beta_5 female + u.$$

(iv) The null hypothesis is H_0: $\delta_1 = 0,\ \delta_2 = 0,\ \delta_3 = 0$, for a total of $q = 3$ restrictions. If n is the sample size, the *df* in the unrestricted model – the denominator *df* in the *F* distribution – is $n - 8$. So we would obtain the critical value from the $F_{q,n-8}$ distribution.

(v) The error term could contain factors, such as family background (including parental history of drug abuse) that could directly affect wages and also be correlated with marijuana usage. We are interested in the effects of a person's drug usage on his or her wage, so we would like to hold other confounding factors fixed. We could try to collect data on relevant background information.

7.10 (i) Yes, simple regression does produce an unbiased estimator of the effect of the voucher program. Because participation was randomized, we can write

$$score = \beta_0 + \beta_1 voucher + u,$$

where *voucher* is independent of u, that is, all other factors affecting *score*. Therefore, the key assumption for unbiasedness of simple regression, Assumption SLR.3, is satisfied.

(ii) No, we do not need to control for background variables. In the equation from part (i), these are factors in the error term, u. But *voucher* was assigned to be independent of all factors, including the listed background variables.

(iii) We should include the background variables to reduce the sampling error of the estimated voucher effect. By pulling background variables out of the error term, we reduce the error variance – perhaps substantially. Further, we can be sure that multicollinearity is not a problem because the key variable of interest, *voucher*, is uncorrelated with all of the added explanatory variables. (This zero correlation will only be approximate in any random sample, but in large samples it should be very small.) The one case where we would not add these variables – or, at least, when there is no benefit from doing so – is when the background variables themselves have no affect on the test score. Given the list of background variables, this seems unlikely in the current application.

C7.2 (i) The estimated equation is

$$\widehat{\log(wage)} = \underset{(0.11)}{5.40} + \underset{(.0063)}{.0654}\, educ + \underset{(.0032)}{.0140}\, exper + \underset{(.0025)}{.0117}\, tenure$$

$$+ \underset{(.039)}{.199}\, married - \underset{(.038)}{.188}\, black - \underset{(.026)}{.091}\, south + \underset{(.027)}{.184}\, urban$$

$n = 935,\ R^2 = .253.$

The coefficient on *black* implies that, at given levels of the other explanatory variables, black men earn about 18.8% less than nonblack men. The t statistic is about –4.95, and so it is very statistically significant.

(ii) The F statistic for joint significance of $exper^2$ and $tenure^2$, with 2 and 925 df, is about 1.49 with p-value $\approx$ .226. Because the p-value is above .20, these quadratics are jointly insignificant at the 20% level.

(iii) We add the interaction $black \cdot educ$ to the equation in part (i). The coefficient on the interaction is about $-.0226$ (se $\approx$ .0202). Therefore, the point estimate is that the return to another year of education is about 2.3 percentage points lower for black men than nonblack men. (The estimated return for nonblack men is about 6.7%.) This is nontrivial if it really reflects differences in the population. But the t statistic is only about 1.12 in absolute value, which is not enough to reject the null hypothesis that the return to education does not depend on race.

(iv) We choose the base group to be single, nonblack. Then we add dummy variables *marrnonblck*, *singblck*, and *marrblck* for the other three groups. The result is

$$\widehat{\log(wage)} = \underset{(0.11)}{5.40} + \underset{(.0063)}{.0655}\, educ + \underset{(.0032)}{.0141}\, exper + \underset{(.0025)}{.0117}\, tenure$$

$$- \underset{(.026)}{.092}\, south + \underset{(.027)}{.184}\, urban + \underset{(.043)}{.189}\, marrnonblck$$

$$- \underset{(.096)}{.241}\, singblck + \underset{(.0560)}{.0094}\, marrblck$$

$n = 935,\ R^2 = .253.$

We obtain the ceteris paribus differential between married blacks and married nonblacks by taking the difference of their coefficients: $.0094 - .189 = -.1796$, or about $-.18$. That is, a married black man earns about 18% less than a comparable, married nonblack man.