Part 1

1) Holding constant the student's age and gender, being in a fraternity or a sorority is associated with an increase of 1.87 days of binge drinking (out of 30 days) (Different wording: The effect on *binge30* [the number of binge-drinking days out of 30 days] of being in a fraternity is 1.87, controlling for the student's age and gender.).

2) The coefficient on *Greek* in regression (1) evidently has omitted variable bias. Students on a sports team do more binge drinking (the coefficient on sports is positive), and being on a sports team seems to be positively correlated with *Greek* (this is also common sense). That is, sports is a positive determinant of *binge30*, and is positively correlated with *Greek*, so when sports is omitted in (1) the coefficient on *Greek* is overstated in (1) and falls when sports is included in the regression.

3) Heteroskedasticity occurs when the variance of the error term depends on one of the regressors. If, for example, the dispersion (or variance) of the number of binge drinking days is larger at a frat than for dorm residents, the variance of the error would depend on Greek.

4) If all Freshmen were 18, all Sophomores 19, all Juniors 20, and all Seniors 21, then *Freshman*, *Sophomore*, *Junior*, *age,* and the "constant" regressor (1) would be perfectly multicollinear:
$18\times Freshman + 19\times Sophomore + 20\times Junior + 21\times(1 - Freshman - Sophomore - Junior) = age$

5) a) All their *X*'s have the same value except for *age*, so
$\Delta binge30 = \hat{\beta}_{age} \times \Delta age = .09\times 2 = .18$ binge drinking days (per 30 days)
b) SE($\Delta \widehat{binge\,30}$) = SE($\hat{\beta}_{age}\times 2$) = 2SE($\hat{\beta}_{age}$) = 2×.10 = .20, so the 95% confidence interval for $\Delta binge30$ is .18 ± 1.96×.20 = [-.21, .57]

6) Thus the hypothesis can be tested using the t-statistic testing $\beta_{age^2} = 0$. That t-statistic is t = -.081/.062 = -1.31, which is less than 1.96 in absolute value, so the hypothesis is not rejected at the 5% significance level.

7) You want to allow the possibility that the effect of sports is nonzero for men and zero for women. This can be achieved by defining the variable $male = 1 - female$ and creating the interaction variable $male\times sports$, then including it in regression (3). The relevant part of the regression thus would be,

$$binge30 = \ldots + \beta_4 sports + \beta_5 male\times sports + \ldots$$

so the effect of sports for women is β_4 and the effect for men is $\beta_4 + \beta_5$. Thus the hypothesis that the effect for women is zero (but not necessarily so for men) can be tested by testing $\beta_4 = 0$. Alternatively, if you use the interaction $female\times sports$ instead of $male\times sports$, you just need to formulate the correct hypothesis.

Part 2

1)

a) False. Regression (5) addresses this by including the interaction between *Greek* and *female*. The effect of *Greek* for men is 2.69 (holding constant the other regressors), while the effect of *Greek* for women is 2.69 – 2.06 = 0.63, far less in real-world terms. The difference is statistically significant at the 1% level ($t = -2.06/.66 = -3.12$), so the hypothesis that binge drinking rates at fraternities and sororities is the same is rejected at the 1% level, with the rate at sororities far less than at fraternities.

b) Depending on what regression you look at, there is some "growing up" effect – lower binge drinking rates for older classes – but the quantitative effect is fairly small. In all regressions (3) – (5), the binge drinking rate is between .35 and .76 days (per 30 days) greater for Freshman than for Seniors, and in all these regressions the rate for freshman exceeds the rate for other classes. On the other hand, in regressions (3) – (5) these differences between classes are not statistically significant (either the coefficient on *Freshman* alone, or jointly with an *F*-statistic testing the coefficients on *Freshman*, *Sophomore*, and *Junior* all being zero), so in those specifications the hypothesis of no difference among classes cannot be rejected.

2) a) Among regressions (1)-(3) we would prefer (3) as it got rid of omitted variable bias due to sport and race variables. Regression (4) is an attempt to model age dependence in non-linear fashion, but the coefficient on quadratic term is insignificant, so we would prefer (3) to it. Finally, (5) allows effect of sororities to be different from effect of fraternities. The difference is significant, so we would choose (5).

b)Taken at face value, the results indicate that fraternity membership is associated with a statistically significant and large (in a real-world sense) increase in the binge drinking rate, holding constant other student characteristics. Based on regression (5), fraternity membership is associated with an increase in the binge drinking rate of nearly 3 days per month on average ($t = 4.80$), holding constant other student characteristics.

In contrast, the results in the table, regression (5) in particular, indicate that sorority membership has much less effect on binge drinking (the hypothesis that it has no effect cannot be tested using only the results provided in the table), increasing binge drinking by only two-thirds of a day per month.

3) One major threat is omitted variable bias, associated with unobserved student characteristics. Suppose students who want a "party life" will drink substantially wherever they reside, and fraternities simply provide an opportunity for these students to live with like-minded fellow partiers. Then the causal effect of being in a fraternity could be small or even zero – Greek is positively correlated with a third variable ("party animal") which is an unobserved individual characteristic and is a determinant of binge30, so the coefficient on Greek in regressions (1) – (5) is biased up.

A second threat arises from these data being self-reported, raising questions about the veracity of the binge30 response. Note that this is not an issue of errors-in-variables bias associated with mismeasurement of the X's – it is plausible that the students accurately reported their age, gender, college year, race, and sports teams. Instead, this is best thought of as omitted variable bias, where the omitted variable is "respondent exaggeration." If frat respondents systematically exaggerated their drinking exploits but other respondents did not, then Greek would be correlated with the omitted variable "respondent exaggeration", so the coefficient on Greek would reflect not a true effect of being in a frat but instead would just measure the fact that frat members exaggerated more than others. Of course, the direction of the bias could go the other way, if frat members tried to hide their binge drinking and systematically underreported, relative to non-frat members.

A third threat is possible sample selection because of the 65% response rate. For sample selection to occur, the sample must be selected by a mechanism that is related to the dependent variable. For example, if heavy binge drinkers were too busy drinking to fill out the survey then there would be sample selection bias (the selection mechanism – busy drinking – is related to the dependent variable – binge drinking).

Part 3

1) False. R^2 always increases when an additional regressor is added whether or not the coefficient is significant.

2) False. This assumption is non-testable. For all samples we have $\sum_{i=1}^{n} x_i \hat{u}_i = 0$

3) False. The correct formula for t-statistic is

$$t = \frac{\hat{\beta}_1 - \hat{\beta}_2}{\sqrt{\operatorname{var}(\hat{\beta}_1) - 2\operatorname{cov}(\hat{\beta}_1, \hat{\beta}_2) + \operatorname{var}(\hat{\beta}_2)}}$$

4) True. F-test take in to account correlation between different t-statistics and is a better way than trying to conduct multiple t-tests, it can deliver a result different from the result of multiple t-tests.

5) False. Coefficient β_3 characterizes the difference of value of a college diploma for women versus value of a college diploma for men.