Sketchy Key and Discussion about Exam 1, Econ 420, Fall 2014

1. 4.597341 is the average female wage, or $E(wage|female)$. 2.556737 is the conditional standard deviation (for female only). We can run a simple regression using female to get 4.597341. The stata command is `reg wage female`, then $\hat{\beta}_0 + \hat{\beta}_1 = 4.597341$. Notice that $\hat{\beta}_1$ alone measures the *difference* between female and male, so is NOT equal to 4.597341. Some students want to use command `reg wage female if female==1`. In theory this is wrong because female will be perfectly correlated with the intercept term if using female only, since there is no variation in the female (and it behaves like a constant).

2. The unconditional mean is $E(y)$; while the conditional mean is $E(y|x)$. The former is a constant, while the latter is a random variable (that depends on x). The former is about the whole population, while the latter is about the subpopulation specified by x (you have two subpopulations if x can only take two values). They are linked via the law of iterated expectation $E(y) = E[E(y|x)]$.

3. $\hat{\beta}_1 = \frac{S_{xy}}{S_x^2} = \frac{4.17645}{7.66749} = 0.54$.

4. The condition or assumption to infer causality from a simple regression is $\texttt{cov}(x, u) = 0$. This holds when ceteris paribus holds. But ceteris paribus can be too strong (ceteris paribus requires u be constant). Doing experiment means we need to randomize educ (we randomly assign values to educ. For example, a person goes to college if and only if his social security number is even). A randomized experiment (or trial) will guarantee $\texttt{cov}(x, u) = 0$.

5. One example of omitted variable is ability. It has positive effect on wage (so $\beta_2 > 0$), and ability is positively correlated with educ ($\texttt{cov}(abiliy, educ) > 0$). In that case, the slope coefficient in the simple regression will overestimate the true marginal effect. Read Table 3.2.

6. You can draw a graph that looks like a facing-down cup (parabola). The left part of that parabola indicates that the wage will go up at *decreasing* rate when exper rises. The optimal exper is given by $\frac{-\beta_1}{2\beta_2} = \frac{-0.2986872}{2(-0.0061397)} = 24.32$. Once the experience exceeds 24.32, the wage can actually go down. Does that make sense?

7. For a *male* worker, a one-year increase in exper is *associated* with 0.66% increase in his wage. For a *female* worker, after a one-year increase in exper, the increase in her

wage is 0.59% *less* than a male worker. You CANNOT say a one-year increase in exper decreases wage by 0.59% for a female worker. Many students do not know how to interpret coefficient when log is involved. Read table 2.3 on page 44. They also have trouble understanding the dummy-variable regression.

8. The unrestricted regression is `reg wage female exper i`; the restricted regression is `reg wage exper`. The null hypothesis of the Chow test is that the female and interaction term both have zero coefficients in the unrestricted regression: $H_0 : \beta_1 = 0, \beta_3 = 0$. We can get RSS from the unrestricted and restricted regression, and then construct the F test. See the lecture note or book for the detailed formula of F test. It is WRONG to say the restricted regression is `reg wage exper if female==1`. That is group-wise regression.

9. Heteroskedasticity means nonconstant conditional variance of the error term $var(u|x) = E(u^2|x) \neq$ `constant`. If you use command `reg y x`, the standard error, t value, p value and confidence interval are all wrong. How about R^2? Also the OLS is no longer BLUE. The WLS or GLS is BLUE.

10. The null hypothesis is that married and unmarried people have the same average wage. We reject this hypothesis since $t = -5.35$ is greater than 1.96 in absolute value. We cannot use this two-sample t test to infer casuality because it does not take into account the effects of other factors. In other words, married and unmarried people earn different wage maybe not because their marital status, but because of other factors.

11. Multicollinearity means regressors are highly correlated with each other. The consequence is big standard error (or imprecise estimate), and small t value.

12. In step 1, regress x_1 onto other regressors, and save residual rhat. In step 2, regress y onto rhat.

13.

$$\begin{aligned}
\hat{\beta} &= \frac{\bar{y}}{\bar{x}} = \frac{n^{-1}\sum y_i}{n^{-1}\sum x_i} && (1)\\
&= \frac{\sum y_i}{\sum x_i} && (2)\\
&= \frac{\sum(\beta x_i + u_i)}{\sum x_i} && (3)\\
&= \beta + \frac{\sum u_i}{\sum x_i} && (4)
\end{aligned}$$

So $E(\hat{\beta}) = \beta$ if we assume $E\left(\frac{\sum u_i}{\sum x_i}\right) = 0$. That holds if the sample is iid, and $E(u|x) = 0$.