Key and discussion, eco311, exam 2, fall 2014

1. In general when $x = 0$ the regression becomes $y = \beta_0 + u$. Assuming $E(u) = 0$, then $\beta_0 = E(y)$. For this problem, *on average* $y = 7.115328$ when $x = 0$. Some students confused the intercept coefficient β_0 with the slope coefficient β_1.

2. The null hypothesis implies that a one-unit increase in x is associated with a one-unit *decrease* in y. The t test is computed as $t = \frac{\hat{\beta}_1 - \beta_1}{\texttt{se}} = \frac{-2.517987-(-1)}{0.3048} = -4.98$. Because the t value is *greater* than 1.96 *in absolute value*, we *reject* the null hypothesis at the 5% level. Some students switched $\hat{\beta}_1$ and β_1; some used the wrong critical value; some drew wrong conclusion; some forgot to interpret the null hypothesis. This is a 10-point question. Detailed answer is expected.

3. The critical value should be 1.645. The 90% confidence interval for β_1 is $\hat{\beta}_1 \pm 1.645 \times se = -2.517987 \pm 1.645 \times 0.3098 = (-3.02, -2.02)$. Loosely speaking, the true value β_1 will be between these two values with 90% probability. Some students used incorrect critical values; some gave me wrong interpretation.

4. $R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{6390.37876}{7222.66439} = 0.12$. So about 12% variation in y can be explained by x. In this case because R^2 is close to zero, x does *not* have much explanatory power.

5. $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x = 7.115328 + (-2.517987)40 = -93.60$

6. For simple regression we have formula $\texttt{var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_i (x_i - \bar{x})^2}$. So decreasing variation in x will lead to smaller denominator, i.e., *increasing* variance and standard error of $\hat{\beta}_1$, and decreasing t value. In practice, we always prefer a sample in which x varies a lot. Intuitively, a lot of variation in x provides a lot of information which can be used to precisely estimate the marginal effect of x on y. If x does not vary, then it is effectively equal to the constant.

7. In general, the error term u represents the unobserved factors that affect y but are excluded in the regression. The residual $\hat{u}$ is the difference between the true and fitted values: $\hat{u} \equiv y - \hat{y}$. So the residual measures the prediction error, or the residual captures the part of y that cannot be explained by x. The regression over-predicts y if the residual is negative. Residual is very useful, for instance, in investment.

8. The two conditions for omitted variable are (A) $\beta_2 \neq 0$; and (B) $\texttt{cov}(x, w) \neq 0$. If both conditions are satisfied, then x in the simple regression will be endogenous, and

OLS estimate will be biased. For this problem we believe VE is an omitted variable. First, common sense is that $\beta_2 > 0$ because students will have higher SAT scores (y goes up) if their parents value education highly. In addition $\texttt{cov}(x, w) = \texttt{cov}(familyincome, VE) > 0$ since the family income tend to be high if parents themselves value education highly and study hard. Jointly these two inequalities imply a *positive* omitted variable bias. See Table 3.2 in the textbook. So the slope coefficient in the simple regression *overestimates* the effect of family income on SAT score. Many students reported the two conditions for omitted variable incorrectly.

9. We need to draw a parabola (cup) facing *down* since $\beta_2 < 0$. As x rises, y first rises at *decreasing* rate. After the turning point, y decreases at *increasing* rate. The quadratic term is necessary if common sense tells us there is boundary for y or there is turning point. Statistically speaking, the quadratic term is necessary if the t value of the quadratic term coefficient is greater than 1.96 in absolute value (or p value is less than 0.05). Many students forgot to discuss the t value, p value or common sense.

10.

$$\begin{aligned} \sum y_i^2 &= \sum (\hat{y}_i + \hat{e}_i)^2 && (1) \\ &= \sum \hat{y}_i^2 + \sum \hat{e}_i^2 + 2\sum \hat{y}_i \hat{e}_i && (2) \\ &= \sum \hat{y}_i^2 + \sum \hat{e}_i^2 && (3) \end{aligned}$$

The last step follows because the first order condition (FOC) of OLS implies that $\sum \hat{y}_i \hat{e}_i = \sum (\hat{\beta}_0 + \hat{\beta}_1 x_i)\hat{e}_i = \hat{\beta}_0 \sum \hat{e}_i + \hat{\beta}_1 \sum x_i \hat{e}_i = \hat{\beta}_0 \times 0 + \hat{\beta}_1 \times 0 = 0$

11. According to the FOC:

$$\begin{aligned} \sum \hat{u}_i = 0 \;&\Rightarrow\; \sum (y_i - \hat{\beta}_0) = 0 && (4) \\ &\Rightarrow\; \sum y_i - n\hat{\beta}_0 = 0 && (5) \\ &\Rightarrow\; \hat{\beta}_0 = \frac{\sum y_i}{n} = \bar{y} && (6) \end{aligned}$$

This result is very important for the next chapter we will study.

12. This is a ten-point question, and detailed discussion is required. Under the null hypothesis $y = \beta_0 + x_1 + u$, so $y - x_1 = \beta_0 + u$. Thus we need to generate a new variable

$z = y - x_1$, and run that strange regression in Q11 using z as the dependent variable. That is how we specify the restricted regression explicitly. The stata codes (not required for the exam) are `gen z = y - x1`; `reg z`. In other words, the restricted regression is *not* `reg y x1` since there is no guarantee the coefficient of x_1 will be equal to one. Nevertheless, if we use z then the restriction of $\beta_1 = 1$ must be imposed. $q = 2$ for this problem. We need to keep the RSS from the restricted and unrestricted regressions. The formula for F test is $F = \frac{(RSS^r - RSS^{ur})/q}{RSS^{ur}/(n-k^{ur}-1)}$. We reject the null hypothesis if the F value is greater than the critical value, or equivalently, if the p value is less than 0.05. Most students didn't know how to specify the restricted regressions. Some used incorrect formula for the F test.

13. This is a multiple level-log regression. The interpretation of the $\hat{\beta}_1$ is *Holding baths and age constant*, increasing area by 1 *percent* is *associated* with rprice increasing by $393.4111 = \frac{39341.11}{100}$ *dollars*. Many students got wrong answers. Please read Table 2.3 in the textbook.

14. The null hypothesis is $H_0 : \beta_3 = 0$. We reject it because the t value = -3.38, greater than 1.96 in absolute value, or p value = 0.000, less than 0.05. Many students specified incorrect null hypothesis.

15. This is a ten-point question, and detailed discussion is required. First, we can use common sense or economic knowledge, both indicate that area may have *non-constant* marginal effect on rprice. So the level-level model considered by Victor may be inappropriate. Second, in terms of statistics, we may run two regressions. One use logarea and the other use area as regressors. Both should use rprice as the dependent variable. The two models are *non-nested*, but have the *same dependent variable*. So we need to select the model that produces higher adjusted R squared.

16. Mathematically, spurious causality arises when $\beta_1 = 0$ (x has no causal effect), but

$$\hat{\beta}_1 \to \beta_1 + \frac{\beta_2 cov(x,w)}{var(x)} = 0 + \frac{\beta_2 cov(x,w)}{var(x)} = \frac{\beta_2 cov(x,w)}{var(x)} \neq 0.$$

Here we assume both conditions of omitted variable hold, see Q8 in this exam. In words, $\hat{\beta}_1$ captures the effect of w instead, despite that x has no causal effect. When designing the Monte Carlo, we need to generate y carefully, by imposing $\beta_1 = 0$. However, because x and w are correlated (by construction), the simple regression of y onto x will yield a statistically significant $\hat{\beta}_1$. That is spurious causality.