

Part I: You have one blue book. That's it! Your answers to these two questions must fit within one blue book. In addition, these questions are worth 40 points, so you should spend no more than 40 minutes on these two questions.

1. We made a number of assumptions in our Classical Regression Model (CRM) that can be succinctly summarized as follows: $u \sim iid N(0, \sigma^2 I_n)$.

- (8) a. In 50 words or less, explain what this rather terse "statement" means. What CRM assumptions are included?

This concise statement of our assumptions (CRMA) means that the CRM disturbances have: a mean of zero (CRMA #3: $E[u] = 0$); a constant variance across all observations (CRMA #4: $\text{Var}(u) = \sigma^2$); and no covariance between different disturbances (CRMA #5: $\text{Cov}[u_i, u_j] = 0$). We are also assuming that these disturbances are normally distributed (CRMA #6). The *iid* part is essentially redundant with the terms in parentheses above – the first *i* stands for *identically distributed*, which means constant variance (homoskedastic). The second *i* stands for *independently distributed* means there is no covariance between different disturbances.

- (20) b. **What possible problems might arise** that would lead to our disturbances not being distributed in this manner? (Don't worry about zero mean part.) Craft a concise explanation using the following outline: (1) **what are the problems that might arise and which CRMA would be violated**; (2) **give an explicit example of each problem**; (3) for each problem, **identify and explain the consequences**. You are expected to provide equations to support your answers, but do not do any derivations here – they will count against you in our grading. We want to read your **explanations**.

We're ignoring problems with the mean of zero, so we need only focus on the *iid* part of the expression:

- (1) If $\text{Cov}(u) \neq \sigma^2 I_n$, then there are two possible problems.

- ❖ The diagonal elements are not the same, σ^2 . This would be a violation of CRMA #4 – our assumption of homoskedasticity. Instead, we now have $\text{Cov}(u) \neq \sigma^2 I_n$, the variances of the disturbances are different at each observation.
- ❖ The off-diagonal elements of $\text{Cov}(u)$ are non-zero. This would occur when two disturbances of different observations are related, they may have either positive or negative covariance. This is most commonly autocorrelation, correlation across disturbances of different time periods. This is a violation of CRMA# 5. When off-diagonal elements are non-zero, we have $\text{Cov}(u_i, u_j) \neq 0$.

- (2) Examples: Give an example of each problem.

- ❖ Heteroskedasticity: Perhaps the changing variance of the disturbances is related to one, or more, of the independent variables. For example, we may find that the dependent variable "wage" increases with the level of education of an individual. If we assume that a constant variance is affected by level of education to give us the non-constant variance, then:

$$\text{Var}(u_i) = \sigma_i^2 = \sigma^2 \cdot Ed_i.$$

- ❖ Autocorrelation: The best examples are time series examples. Suppose our dependent variable Y includes a disturbance that is related to the previous period's disturbance. Then, we have our typical first-order autocorrelation model: $u_t = \rho u_{t-1} + \varepsilon_t$.

(3) We can explain the consequences of these two problems in general as they both have similar effects on the properties of the OLS parameter estimators.

- ❖ First, the OLS estimators are, in matrix form: $b = (X'X)^{-1}X'Y$. The OLS estimators are unbiased. Recall that we used CRMA # 1-3 to prove that the OLS estimators are unbiased, and these assumptions are not affected by problems in the covariance matrix of the disturbances (remember, we are assuming that the assumption $E[u] = 0$ still holds.)
- ❖ The consequences of these two problems are found in the covariance matrix for the OLS estimators. OLS assumes that all CRMA hold. Under these assumptions the true covariance matrix of the estimators is: $Cov(b) = \sigma^2(X'X)^{-1}$. But **this is wrong, or biased**, because one or more of the assumptions that would allow us to use $Cov(u) = \sigma^2I_n$ in deriving $Cov(b)$ do not hold. Thus, we would use the general covariance matrix, $Cov(u) = W$ in deriving the covariance matrix of the OLS estimators. The true covariance matrix for the OLS estimators is:
$$Cov(b) = (X'X)^{-1}X'W X(X'X)^{-1}.$$
All the software packages will give you the wrong covariance matrix when you use OLS because they are making the wrong assumptions. Thus, all standard errors will be wrong and all calculated t-values will be wrong.
- ❖ The consequence is that you cannot do inference with OLS when either of these assumptions (CRMA# 4 or CRMA# 5) are violated. All inferences, confidence intervals and hypothesis tests, would be wrong. And, because we don't know the direction of the "bias," we will not know if we are over estimating or under estimating the standard errors.

(12) 2. I submitted an article for consideration to a journal. One of my reviewers included the following statement: "...In addition, these time-series data are likely fraught with multicollinearity. Thus, you can not trust the estimates." How would you respond to this chap? Do you think he was on target? Write a concise response that discusses the consequences of multicollinearity and assure him that you indeed do not have a serious problem. (You're not getting a printout for this question; just pretend you have information from your analyses to use in answering the question.)

I really did get this as a comment on a journal article submission. I recall going to Cleve Willis, who was our department chair, and saying something like: "... this guy's a jerk." Cleve said something like: "That's possible, but if you put that in your review, you won't be getting the article published." The following would be a reasonable way to reply:

- ❖ The statement that "... you can not trust the estimates..." could be based on the result that the denominator of the estimators may tend toward zero under severe multicollinearity. Thus the estimates could become very large. In addition, when you add or delete variables, omitted effects would cause specification bias in included effects that might cause sign changes. But, let's make the assumption that the model is properly specified. We would want to assure the reviewer of that and thank him for reminding us to point out this important feature of the research.
- ❖ The true consequences of multicollinearity are that the calculated t-statistics will be deflated or depressed. As a result, the **likelihood of making a Type II Error under our typical test conditions will be high**. To alleviate fears of these problems, you can **include the variance inflation factors** for each independent variable. You could point out that the variance inflation factors are all less than the "rule of thumb" and that there really are no problems with multicollinearity. You could also include a matrix of correlation coefficients. Finally, in the case where some VIFs were above the threshold of 10, you could point out that the consequences are clear and that you've chosen to modify your level of significance. Once again, thank the reviewer for pointing out your omission of this additional information. (Brownie points – you'll surely get this baby accepted!)
- ❖ For the case in question, the model was a time-series model, but all prices were relative prices eliminating problems with multicollinearity.

Part II: Applied questions – use the SAS output provided to answer these questions and write all answers in the space provided on these pages.

I had my RA run some regressions on 2004 New England wages for individuals age 18 and over who worked for private firms. She does careful work, so I generally trust that the analyses she’s provided me are on target. But, there are some additional questions that I had after she did her work and headed off to Bermuda for the weekend.

Use the first, combined male and female wage model to answer the following questions.

- (5) 1. Interpret the estimated effect of education on New England wages.

This is a proportionate change in wage given a one year change in level of education:

$$\frac{\partial \ln \widehat{wage}}{\partial Ed} = \frac{\frac{\partial \widehat{wage}}{\widehat{wage}}}{\partial Ed} = 0.07302.$$

So, a 1-year increase in education will result in a 7.3% increase in wage, holding constant the level of experience. Because this was a proportionate change (see numerator) we multiply by 100 to get a percentage change. (None of the dummy variables interact with the level of education. There is the same effect for males and females, union and non-union, and those with and without a pension plan.)

- (10) 2. Identify the estimated coefficients that are statistically significant in this first “pooled” model. Comment – how did you arrive at these conclusions?

- ❖ With 748 degrees of freedom, I used the value from the t-table of $\alpha = 0.10$ and $df = 1000$. Thus my critical t-values were 1.282 or -1.282 (virtually z-values), depending upon the appropriate form of the alternative hypothesis.
- ❖ I expect positive coefficients for the variables *ed*, *exp*, *m*, *expm*, *unioncov*, *expunion*, *munion* and *pension*. Of these variables, *ed*, *exp*, *expm* and *pension* were statistically greater than zero. This is true because the calculated t-values are greater than the critical t-value.
- ❖ I expected a quadratic effect of education on wage and that this quadratic effect would be “hill-shaped.” Thus, I expect a negative sign for the variable *expsq*. This variable was also statistically less than zero.

- (8) 3. Derive the expression for the partial effect of experience on wages. Explain any “peculiarities” of this expression so that we can interpret the partial effects properly.

The partial effect is:

$$\frac{\partial \ln \widehat{wage}_i}{\partial Ed_i} = 0.03096 - (2)(0.00056) exp_i + 0.0064 m_i - 0.00252 union_i$$

This partial effect changes with experience, and it changes depending upon gender ($m = 1$ or $m = 0$) as well as whether or not the individual is covered by a union contract ($union = 1$ vs $union = 0$). Thus, we can get a set of four different regimes for these partial effects:

All these partial effects would then change with the level of experience.

	Union	Non-Union
Male	(1,1)	(1,0)
Female	(0,1)	(0,0)

- (4) 4. The variable “pension” is a binary variable that indicates whether the individual is covered by a pension plan provided by their employer. Please interpret the estimated coefficient for “pension.”

Because the dependent variable is in natural log form, we need to make a correction. The most straightforward correction is the following:

$$\hat{g} = (\exp^{\hat{\delta}_3}) - 1 = 0.17999.$$

Interpretation: Jobs with a pension plan pay about 18% higher wages than jobs without a pension plan.

- (4) 5. How well does this model fit the data? Explain.

The model explains about 30% of the variation in $\ln wages$. This is measured by $R^2 = 0.2994$.

- (5) 6. Is the regression model statistically significant? Explain.

This the F-test of the regression model. The null and alternative hypotheses are:

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \delta_M = \gamma_{expM} = \delta_U = \gamma_{expU} = \delta_{MU} = \delta_P = 0$$
$$H_a : \text{at least one is not zero.}$$

The calculated value for F is: $F_{calc} = 35.52$. The prob-value for this calculate F-value is < 0.0001 . Clearly, this calculated F-value will exceed the critical F-values for any reasonable choice for the level of significance. This model is statistically significant. It does explain a statistically important portion of the variation in the dependent variable $\ln wage$.

- (3) 7. She marked the final two regressions as “male” and “female.” I’m worried that she may have marked these incorrectly. Is there any way to tell? Explain.

This is easy, and it actually is what happened after I ran the regression and shuffled them around. We know the definition of the dummy variable, m . The mean of m is the proportion of males. If we then multiply the mean of m by the number of observations, we get the number of males. Then just match that with the number of observations used in the regression.

Note, if you assumed that the males had a higher intercept, then you’re essentially completing a hypothesis without using the data to test. However, we do know from the pooled model that the male intercept should be about \$1.23.

(8) 8. I wanted to test whether male and female wages are different. She says it's no problem with the results I have. Please complete this test for me.

(i) State the appropriate null and alternative hypotheses.

$$H_0 : \delta_M = \gamma_{expM} = \delta_{MU} = 0$$

H_a : at least one is not zero.

(ii) Calculate the appropriate test statistic.

The calculated F-statistic is given: $F_{calc} = 13.04$.

(iii) Complete the test. What do you conclude?

$F_{(\alpha, n_1, n_2)} = F_{(0.05, 3, 748)} \approx 2.64$. $F_{calc} > F_{(\alpha, n_1, n_2)}$, so we reject the null hypothesis that male and female wages are the same.

We conclude that male and female wages are statistically different.

(8) 9. Okay, now that we've done all that, we should decide whether we should have used the results for the pooled male/female wage model. We should test for heteroskedasticity and she said these results could be used to do that. How?

(i) What test can be applied using these results? What form of heteroskedasticity could be diagnosed by this test? Explain briefly, then state the null and alternative hypotheses for the hypothesis test.

This is a good place for a Goldfeld-Quandt test: $H_0 : \sigma_M^2 = \sigma_F^2$; $H_a : \sigma_M^2 \neq \sigma_F^2$.

We expect "group-wise" heteroskedasticity; that is the variance for males may be different than the variance for females. Our null hypothesis is that we have homoskedasticity, the alternative is that there is heteroskedasticity.

(ii) Calculate the appropriate test statistic.

The F-statistic is calculated as a ratio of the two estimated variances. Always put the largest on top, in this case the variance for the female wage regression:

$$F_{calc} = \frac{\hat{\sigma}_F^2}{\hat{\sigma}_M^2} = \frac{0.15287}{0.14923} = 1.024.$$

(iii) Complete the test. What do you conclude?

The degrees of freedom are 368 and 376 for the numerator and denominator, respectively. Those values are not in the table. Anything close is fine, or you could have used the table I made up in Excel. The accurate F-value was:

$$F_{(\alpha, n_1, n_2)} = F_{(0.05, 368, 376)} \approx 1.186.$$

Thus, we fail to reject the null hypothesis that these wage data are homoskedastic. We would conclude there is no problem with group-wise heteroskedasticity.

- (5) Given the results of your hypothesis tests, which regression results would you use in drawing inferences about New England wages? Why? Explain.

You can go either way on this one, but you've got to have good reasons.

You could argue that with not group-wise heteroskedasticity, the joint model is the one to use. It has the advantage of providing parameter estimates that measure differences between male and female wages and the effects of experience. The calculated t-statistics for these variables also provide convenient tests of differences as well.

Or you could argue that the F-test above suggested that male and female wages are different. Therefore, it would be best to use the two separate regressions, which allow complete parameter differences across the two groups.