Econometrics -- Final Exam (Sample)

1) The sample regression line estimated by OLS
A) has an intercept that is equal to zero.
B) is the same as the population regression line.
C) cannot have negative and positive slopes.
D) is the line that minimizes the sum of squared prediction mistakes.

2) Under imperfect multicollinearity
A) the OLS estimator cannot be computed.
B) two or more of the regressors are highly correlated.
C) the OLS estimator is biased even in samples of $n > 100$.
D) the error terms are highly, but not perfectly, correlated.

3) Consider the multiple regression model with two regressors X_1 and X_2, where both variables are determinants of the dependent variable. When omitting X_2 from the regression, then there will be omitted variable bias for $\widehat{\beta 1}$

A) if X_1 and X_2 are correlated
B) always
C) if X_2 is measured in percentages
D) if X_2 is a dummy variable

4) The OLS estimator for the multiple regression model in matrix form is
A) $(\boldsymbol{X'X})^{-1}\boldsymbol{X'Y}$
B) $\boldsymbol{X}(\boldsymbol{X'X})^{-1}\boldsymbol{X'} - \boldsymbol{P}_X$
C) $(\boldsymbol{X'X})^{-1}\boldsymbol{X'U}$
D) $(\boldsymbol{X\Omega}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X\Omega}^{-1}\boldsymbol{Y}$

5) The following linear hypothesis can be tested using the F-test with the exception of
A) $\beta_2 = 0$.
B) $\beta_1 + \beta_2 = 1$ and $\beta_3 = -2\beta_4$.
C) $\beta_0 = \beta_1$ and $\beta_1 = 0$.
D) $\beta_3 = \beta_1 + \beta_4\beta_5$

6) The overall regression F-statistic tests the null hypothesis that
A) all slope coefficients are zero.
B) all slope coefficients and the intercept are zero.
C) the intercept in the regression and at least one, but not all, of the slope coefficients is zero.
D) the slope coefficient of the variable of interest is zero, but that the other slope coefficients are not.

7) Let $SSR_{unrestricted}$ and $SSR_{restricted}$ be 56.8 and 63.4 respectively. The difference between

the unrestricted and the restricted model is that you have imposed two restrictions. There are 200 observations, and 3 regressors including the intercept. The homoscedasticity-only *F*-statistic in this case is

A) 2.89

B) 23.17

C) 11.45

D) 5.61

8) A 95% confidence set for two or more coefficients is a set that contains

A) the sample values of these coefficients in 95% of randomly drawn samples.

B) integer values only.

C) the same values as the 95% confidence intervals constructed for the coefficients.

D) the population values of these coefficients in 95% of randomly drawn samples.

9) When your multiple regression function includes a single omitted variable regressor, then

A) use a two-sided alternative hypothesis to check the influence of all included variables.

B) the estimator for your included regressors will be biased if at least one of the included variables is correlated with the omitted variable.

C) the estimator for your included regressors will always be biased.

D) lower the critical value to 1.645 from 1.96 in a two-sided alternative hypothesis to test the significance of the coefficients of the included variables.

10) Let there be q joint hypothesis to be tested. Then the dimension of $\boldsymbol{r}$ in the expression $\boldsymbol{R\beta} = \boldsymbol{r}$ is

A) $q \times 1$.

B) $q \times (k+1)$.

C) $(k+1) \times 1$.

D) q.

11) The interpretation of the slope coefficient in the model $Y_i = \beta_0 + \beta_1 \ln(X_i) + u_i$ is as follows:

A) a 1% change in X is associated with a β_1 % change in Y.

B) a 1% change in X is associated with a change in Y of 0.01 β_1.

C) a change in X by one unit is associated with a β_1 100% change in Y.

D) a change in X by one unit is associated with a β_1 change in Y.

12) Including an interaction term between two independent variables, X_1 and X_2, allows for the following except:

A) the interaction term lets the effect on Y of a change in X_1 depend on the value of X_2.

B) the interaction term coefficient is the effect of a unit increase in X_1 and X_2 above and beyond the sum of the individual effects of a unit increase in the two variables alone.

C) the interaction term coefficient is the effect of a unit increase in $\sqrt{(X_1 \times X_2)}$.

D) the interaction term lets the effect on Y of a change in X_2 depend on the value of X_1.

13) For the polynomial regression model,
A) you need new estimation techniques since the OLS assumptions do not apply any longer.
B) the techniques for estimation and inference developed for multiple regression can be applied.
C) you can still use OLS estimation techniques, but the *t*-statistics do not have an asymptotic normal distribution.
D) the critical values from the normal distribution have to be changed to 1.96^2, 1.96^3, etc.

14) In the regression model $Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + \beta_3 (X_i \times D_i) + u_i$, where X is a continuous variable and D is a binary variable, β_3
A) is the difference in means in Y between the two categories.
B) indicates the difference in the intercepts of the two regression lines.
C) is usually positive.
D) indicates the difference in the slopes of the two regression lines.

15) Assume that you had estimated the following quadratic regression model

$\widehat{TestScore} = 607.3 + 3.85\ Income - 0.0423\ Income^2$. If income increased from 10 to 11 ($10,000 to $11,000), then the predicted effect on testscores would be
A) 3.85
B) 4.74
C) Cannot be calculated because the function is non-linear
D) 2.96

16) To test whether or not the population regression function is linear rather than a polynomial of order *r*,
A) check whether the regression R^2 for the polynomial regression is higher than that of the linear regression.
B) compare the *TSS* from both regressions.
C) look at the pattern of the coefficients: if they change from positive to negative to positive, etc., then the polynomial regression should be used.
D) use the test of (*r*-1) restrictions using the *F*-statistic.

17) The components of internal validity are
A) a large sample, and BLUE property of the estimator.
B) a regression R^2 above 0.75 and serially uncorrelated errors.
C) unbiasedness and consistency of the estimator, and desired significance level of hypothesis testing.
D) nonstochastic explanatory variables, and prediction intervals close to the sample mean.

18) Simultaneous causality
A) means you must run a second regression of X on Y.
B) leads to correlation between the regressor and the error term.
C) means that a third variable affects both Y and X.

D) cannot be established since regression analysis only detects correlation between variables.

19) Applying the analysis from the California test scores to another U.S. state is an example of looking for
A) simultaneous causality bias.
B) external validity.
C) sample selection bias.
D) internal validity.

20) A possible solution to errors-in-variables bias is to
A) use log-log specifications.
B) choose different functional forms.
C) use the square root of that variable since the error becomes smaller.
D) mitigate the problem through instrumental variables regression.

Long Questions

21) You have estimated an earnings function, where you regressed the log of earnings on a set of continuous explanatory variables (in levels) and two binary variables, one for gender and the other for marital status. One of the explanatory variables is education.
(a) Interpret the education coefficient.
(b) Next, specify the binary variables and an equation, where the default is a single male, without allowing for interaction between marital status and gender. Indicate the coefficients that measure the effect of a single male, single female, married male, and married female.
(c) Finally allow for an interaction between the gender and marital status binary variables. Repeat the exercise of writing down the various effects based on the female/male and single/married status. Why is the latter approach more general than the former?

22) Sports economics typically looks at winning percentages of sports teams as one of various outputs, and estimates production functions by analyzing the relationship between the winning percentage and inputs. In Major League Baseball (MLB), the determinants of winning are quality pitching and batting. All 30 MLB teams for the 1999 season. Pitching quality is approximated by "Team Earned Run Average" (ERA), and hitting quality by "On Base Plus Slugging Percentage" (OPS).

Summary of the Distribution of Winning Percentage, On Base Plus Slugging Percentage, and Team Earned Run Average for MLB in 1999

	Average	**Standard deviation**	**Percentile**						
			10%	**25%**	**40%**	**50% (median)**	**60%**	**75%**	**90%**
Team ERA	4.71	0.53	3.84	4.35	4.72	4.78	4.91	5.06	5.25

OPS	0.778	0.034	0.720	0.754	0.769	0.780	0.790	0.798	0.820
Winning Percentage	0.50	0.08	0.40	0.43	0.46	0.48	0.49	0.59	0.60

Your regression output is:

$$\widehat{Winpct} = -0.19 - 0.099 \times teamera + 1.490 \times ops,\ R^2=0.92,\ SER = 0.02.$$
$$(0.08)\quad (0.008)\qquad (0.126)$$

(a) Interpret the regression. Are the results statistically significant and important?
(b) There are two leagues in MLB, the American League (AL) and the National League (NL). One major difference is that the pitcher in the AL does not have to bat. Instead there is a "designated hitter" in the hitting line-up. You are concerned that, as a result, there is a different effect of pitching and hitting in the AL from the NL. To test this hypothesis, you allow the AL regression to have a different intercept and different slopes from the NL regression. You therefore create a binary variable for the American League (*DAL*) and estimate the following specification:

$$\widehat{Winpct} = -0.29 + 0.10 \times DAL - 0.100 \times teamera + 0.008 \times (DAL \times teamera)$$
$$(0.12)\quad (0.24)\qquad (0.008)\qquad (0.018)$$

$$+ 1.622 * ops - 0.187 * (DAL \times ops),\ R^2=0.92,\ SER = 0.02.$$
$$(0.163)\qquad (0.160)$$

What is the regression for winning percentage in the AL and NL? Next, calculate the t-statistics and say something about the statistical significance of the AL variables. Since you have allowed all slopes and the intercept to vary between the two leagues, what would the results imply if all coefficients involving DAL were statistically significant?
(c) You remember that sequentially testing the significance of slope coefficients is not the same as testing for their significance simultaneously. Hence you ask your regression package to calculate the F-statistic that all three coefficients involving the binary variable for the AL are zero. Your regression package gives a value of 0.35. Looking at the critical value from you F-table, can you reject the null hypothesis at the 1% level? Should you worry about the small sample size?

Table 1 Critical Values for the $F_{m,\infty}$ Distribution

	Significance level		
Degrees of freedom	10%	5%	1%
1	2.706	3.841	6.635
2	2.303	2.996	4.605
3	2.084	2.605	3.782
4	1.945	2.372	3.319
5	1.847	2.214	3.017
6	1.774	2.099	2.802
7	1.717	2.010	2.639
8	1.670	1.938	2.511
9	1.632	1.880	2.407
10	1.599	1.831	2.321

This table contains the 90^{th}, 95^{th}, and 99^{th} percentiles of the $F_{m,\infty}$ distribution. These serve as critical values for tests with significance levels of 10%, 5%, and 1%.

Answer:

1) D
2) B
3) A
4) A
5) D
6) A
7) C
8) D
9) B
10) A
11) B
12) C
13) B
14) D
15) D
16) D
17) C
18) B
19) B
20) D

21)

(a) The coefficient on education gives you the return to education, i.e., if education increased by one year, then by how many percent do earnings increase?

(b) Let *DGender* equal one if the individual is a female, and be zero otherwise. *DMarried* takes on a value of one if the individual is married and is zero otherwise. The regression is

$\widehat{\ln Earn} = \hat{\beta}_0 + \hat{\beta}_1 DGender + \hat{\beta}_2 DMarried + ...$

Single male: $\hat{\beta}_0$; single female: $\hat{\beta}_0 + \hat{\beta}_1$; married male: $\hat{\beta}_0 + \hat{\beta}_2$; married female: $\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2$.

(c) $\widehat{\ln Earn} = \hat{\beta}_0 + \hat{\beta}_1 DGender + \hat{\beta}_2 DMarried + \hat{\beta}_3 (DGender \times DMarried + ...$

Single male: $\hat{\beta}_0$; single female: $\hat{\beta}_0 + \hat{\beta}_1$; married male: $\hat{\beta}_0 + \hat{\beta}_2$; married female: $\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}_3$. This approach is more general because it allows the effect of being married and female to be different from being married and male. In (b), both females and males were faced with identical effects from being married, $\hat{\beta}_2$. In (c), this effect differs due to the additional coefficient $\hat{\beta}_3$.

22)

(a) Lowering the team ERA by one results in a winning percentage increase of roughly ten percent. Increasing the OPS by 0.1 generates a higher winning percentage of approximately 15 percent. The regression explains 92 percent of the variation in winning percentages. Both slope coefficients are statistically significant, and given the small differences in winning percentage, they are also important.

(b) NL: $\widehat{Winpct} = -0.29 - 0.100 \times teamera + 1.622 \times ops$.

AL : $\widehat{Winpct} = -0.19 - 0.092 \times teamera + 1.435 \times ops$.

The *t*-statistics for all variables involving *DAL* are, in order of appearance in the above regression, 0.42, 0.44, and –1.17. None of the coefficients is statistically significant individually. If these were statistically significant, then this would indicate that the coefficients vary between the two leagues. Hence it would suggest that the introduction of the designated hitter might have changed the relationship.

(c) The critical value of the *F*-statistic is 3.78 at the 1% level, and hence you cannot reject the null hypothesis, that all three coefficients are zero. However, the *F*-statistic is not really distributed as $F_{3,\infty}$, and, as a result, inference is problematic here.