Sample exam

You have 80 minutes to complete this setup. The exam is worth 60 points in total, exact amount of points is indicated for each exercise.

1. **(10 points)** Describe the heteroskedasticity problem: explain briefly what it is and how it affects the estimation. Give the name of at least one of the tests for heteroskedasticity and state its null hypothesis. Provide at least one solution of the heteroskedasticity problem.

2. **(6 points)** Describe the two properties a valid instrumental variable must satisfy.

3. **(5 points)** Decide if the following claim is true or false (and explain why): "I run a regression of y on x and I save the residuals $e = y - \widehat{y}$. If I find that $Cov(x, e) = 0$, I have the right to conclude that the variable x was exogenous in my regression."

4. Suppose a sample of adults is classified into groups 1, 2 and 3 on the basis of whether their education stopped at the end of elementary school, high school, or university, respectively. The relationship

$$y = \beta_1 + \beta_2 D_2 + \beta_3 D_3 + \varepsilon$$

   is specified, where y is income, $D_i = 1$ for those in group i and zero for all others.

   (a) **(3 points)** Explain why D_1 is not included in the regression.

   (b) **(4 points)** In terms of the parameters of the model, what is the expected income of people whose education stopped at the end of university? What is the expected income of people whose education stopped at the end of elementary school?

   (c) **(6 points)** Suppose some respondents who only finished the elementary school were embarrassed about their lack of education and lied, claiming that they graduated from high school. What would be the impact of this lie on the estimates of β_2 and β_3?

5. **(26 points)** You have data for 732 student-athletes from a large university for fall semester. Your primary question of interest is this: Do athletes perform more poorly in school during the semester when their sport is in season?

   (a) You run a regression with the following variables:

trmgpa	...	the student's GPA (grade point average) for the semester
season	...	dummy, equal to 1 if the student's sport is in season that semester, 0 otherwise
hsrank	...	the student's performance at high school, measured as the rank among his/her classmates
crsgpa	...	the course GPA (average GPA over all students taking the course) for the semester

   You obtain the following result:

Source	SS	df	MS	Number of obs =	732
				F(3, 728) =	48.49
Model	69.996157	3	23.3320523	Prob > F =	0.0000
Residual	350.300799	728	.481182416	R-squared =	0.1665
				Adj R-squared =	0.1631
Total	420.296956	731	.574961636	Root MSE =	.69367

trmgpa	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
season	-.1091708	.0546164	-2.00	0.046	-.2163952	-.0019463
hsrank	-.0021305	.0002344	-9.09	0.000	-.0025907	-.0016703
crsgpa	.9246755	.1165595	7.93	0.000	.6958426	1.153508
_cons	-.0829782	.327399	-0.25	0.800	-.725737	.5597806

   i. Interpret the coefficient on *season*. Is it significant at 95% confidence level?

   ii. Test if the coefficient on *crspga* is significantly different from 1 at 95% confidence level[1]. Interpret your finding.

   iii. Explain what *hsrank* controls for in the regression. (*Hint*: the lower *hsrank*, the better the high school performance of the student).

[1] for critical values, see Appendix on p.3

(b) Most of the athletes who play their sport in the fall are football players. You suppose the academic performance of football players differ systematically from those of other athletes.

You include in you regression a new variable *football*, which is a dummy equal to one if the student is football player, zero otherwise. You get the following result:

Source	SS	df	MS
Model	99.9162975	4	24.9790744
Residual	320.380658	727	.440688664
Total	420.296956	731	.574961636

Number of obs	= 732
F(4, 727)	= 56.68
Prob > F	= 0.0000
R-squared	= 0.2377
Adj R-squared	= 0.2335
Root MSE	= .66384

trmgpa	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
season	.0018219	.0539756	0.03	0.973	-.1041449	.1077886
hsrank	-.001966	.0002252	-8.73	0.000	-.0024081	-.0015238
crsgpa	.8978952	.1115946	8.05	0.000	.6788091	1.116981
football	-.4479337	.0543623	-8.24	0.000	-.5546595	-.3412078
_cons	.1937855	.3151154	0.61	0.539	-.4248594	.8124303

Is the coefficient on *season* significant at 95% confidence level now? Does this confirm there was a bias in part (a)? If yes, explain how this bias was created and what was its sign (intuitive explanation is sufficient).

Appendix:
Extract from statistical table of Student t-distribution (area under right-hand tail)

d.f.	**0.05**	**0.025**	**0.01**
40	1.684	2.021	2.423
60	1.671	2.000	2.390
120	1.658	1.980	2.358
∞	1.645	1.960	2.326