Final Exam

EC031-S26

Note: Please make sure to read all the instructions carefully, especially as it pertains to the Stata portion of the exam.

Choose and answer 4 out of 5 questions. You will not receive extra credit for answering all questions.

The exam is out of 100 points. By default, all math-involved questions require showing your work, unless otherwise stated. Simply writing down the right answer is not enough for full credit.

For "explain your answer" questions, the length of the answer is not as important as being able to clearly explain your thinking. As a general rule, 3-5 sentences should suffice.

The amount of space that I give you is not indicative of how much is needed to answer the question. Clearly mark the questions that you choose.

Note: The last page of the exam is an F-distribution table.

The Stata portion of the exam will be given in the form of a Moodle quiz.

You cannot start the Stata portion of the exam, until you have completed the written portion of the exam.

Question 1 [25 points]

You are estimating the following model on a sample of 100 observations:

$$\text{consumption}_i = \beta_0 + \beta_1 \cdot \text{income}_i + \epsilon_i$$

To test for heteroskedasticity, you divide the sample into two groups:
- Group 1: Lowest 1/3 of observations based on `income`
- Group 2: Highest 1/3 of observations based on `income`
(You drop the middle third.)

You then estimate the model separately on both groups. Below is the output from the two regressions:

Group 1: Lowest 1/3 of observations based on `income`

```
      Source |       SS           df       MS      Number of obs   =        34
-------------+----------------------------------   F(1, 32)        =      0.09
       Model |  138.450031         1  138.450031   Prob > F        =    0.7678
    Residual |  49952.4832        32   1561.0151   R-squared       =    0.0028
-------------+----------------------------------   Adj R-squared   =   -0.0284
```

```
       Total |  50090.9332        33  1517.90707   Root MSE        =     39.51

------------------------------------------------------------------------------
 consumption | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
      income |   .1600444   .5373998     0.30   0.768    -.9346031    1.254692
       _cons |   66.08665   40.42751     1.63   0.112     -16.2615    148.4348
------------------------------------------------------------------------------
```

Group 2: Highest 1/3 of observations based on `income`

```
      Source |       SS           df       MS      Number of obs   =        33
-------------+----------------------------------   F(1, 31)        =      0.04
       Model |  210.439315         1  210.439315   Prob > F        =    0.8384
    Residual |  154201.659        31  4974.24707   R-squared       =    0.0014
-------------+----------------------------------   Adj R-squared   =   -0.0309
       Total |  154412.098        32  4825.37808   Root MSE        =    70.528

------------------------------------------------------------------------------
 consumption | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
      income |    .150965   .7339671     0.21   0.838    -1.345971    1.647901
       _cons |    116.163   98.97162     1.17   0.249    -85.69093     318.017
------------------------------------------------------------------------------
```

a. Compute the Goldfeld-Quandt F-statistic.

b. Based on your result and a 10% level of Type-I error, what do you conclude about the presence of heteroskedasticity? (You only need to guesstimate the critical value.)

c. What does this imply about OLS and its assumptions?

d. In order to account for the heteroskedasticity, you decide to use robust standard errors. Below are the two regression tables for the *full sample* of 100 observations. The first shows the regression results assuming homoskedasticity, and the second shows the regression results with robust standard errors.

How did the standard errors change? What does this imply about the significance of the coefficients now? If you had not calculated robust robust standard errors, would you be over- or under-estimating the significance of the coefficients?

Full Sample, Homoskedasticity Assumed

```
      Source |       SS           df       MS      Number of obs   =       100
-------------+----------------------------------   F(1, 98)        =     13.80
       Model |  52562.0813         1  52562.0813   Prob > F        =    0.0003
    Residual |  373251.643        98  3808.69024   R-squared       =    0.1234
-------------+----------------------------------   Adj R-squared   =    0.1145
       Total |  425813.725        99  4301.14874   Root MSE        =    61.715

------------------------------------------------------------------------------
 consumption | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
      income |   .8338093   .2244493     3.71   0.000     .3883969    1.279222
       _cons |   26.49095   24.14265     1.10   0.275    -21.41936    74.40125
------------------------------------------------------------------------------
```

Full Sample, Robust Standard Errors

```
Linear regression                               Number of obs     =       100
                                                F(1, 98)          =     15.23
                                                Prob > F          =    0.0002
                                                R-squared         =    0.1234
                                                Root MSE          =    61.715

------------------------------------------------------------------------------
             |               Robust
 consumption | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
      income |   .8338093   .2136697     3.90   0.000     .4097887     1.25783
       _cons |   26.49095   20.21286     1.31   0.193    -13.62082    66.60271
------------------------------------------------------------------------------
```

Question 2 [25 points]

Suppose we would like to see the effect of poverty on academic performance in California high schools. To capture poverty, we decide to use free lunch eligibility and to capture performance, we decide to use the API (academic performance index) which is collected from all California high schools. We collect a representative sample and run a regression to the following population model:

$$API_i = \beta_0 + \beta_1 FLE_i + \varepsilon_i$$

We are concerned that parental education is not included in the regression, thus creating correlation between FLE_i and ε_i.

a. Do you think that our regression gives a causal impact of free lunch eligibility on academic performance? Why or why not? What would be the direction of the bias?

Let's say that we don't have any observations on parental education as parents refused to give that information. We decide to think about an instrument that we could use instead. For each of the following instruments for FLE_i, explain why or why not they would be good instruments. Form you answer in terms of relevance and the exclusion restriction (for the exclusion restriction, you can assume that the only source of endogeneity is parental education):

b. The average income of each county in California

c. The number of students who own cars.

d. A natural disaster that randomly hit some counties in California that are in the sample.

Question 3 [25 points]

In class we learned that within the potential outcomes framework, the simple difference in means (SDM) between the treatment and controls groups can be decomposed into the following:

$$\begin{aligned} E[Y \mid D=1] - E[Y \mid D=0] = E[Y^1] - E[Y^0] \quad & \text{Term 1} \\ + E[Y^0 \mid D=1] - E[Y^0 \mid D=0] \quad & \text{Term 2} \\ + (1-\pi)(ATT - ATU) \quad & \text{Term 3} \end{aligned}$$

where π is the proportion of the treated group.

a. What are ATT and ATU? Why aren't they equal?

b. What is Term 1 in the above decomposition?

c. What is Term 2 in the above decomposition?

d. What is Term 3 in the above decomposition?

Below is a table of potential outcomes with a sample of patients that received surgery as a treatment and those that didn't. The outcome is defined as the number of years that patients lived post-surgery.

A "perfect, benevolent doctor" sorted patients into receiving or not receiving surgery based on observing each patient's potential outcomes and then choosing the treatment that would yield the best outcome for them.

The **potential outcomes table** has the following columns:

- Y^1: The potential outcome if receiving surgery
- Y^0: The potential outcome if not receiving surgery
- D: The actual treatment received (1 = surgery, 0 = no surgery)
- Y: The actual observed outcome

Patient	Surgery (Y^1)	No Surgery (Y^0)	D	Y
1	8	1	1	8
2	5	6	0	6
3	5	1	1	5
4	7	8	0	8
5	5	2	1	5
6	3	4	0	4
7	6	4	1	6

e. What variables are actually observed in the table? What variables are not usually observed? Explain why we can never observe both potential outcomes for the same individual simultaneously.

f. What is the value of the SDM?

g. What is the value of Term 1?

h. What is the value of Term 2?

i. What is the value of Term 3?

j. Verify that the SDM is equal to the sum of the three terms.

Question 4 [25 points]

You're analyzing the effect of vaccination (V) on infection rate (I), where health status (H) affects both V and I, and hospital visits (HV) are affected by both V and I.

a. Draw the DAG.

b. Write the linear regression specification for each variable in the DAG if the DAG implies that the variable is a function of the other variables.

c. Identify a confounder and a collider. Explain why they are a confounder or a collider.

d. Explain what would happen if you controlled for hospital visits in your regression.

e. What is the backdoor path between V and I, and how would you block it?

f. If this DAG were true, what would be the best way to estimate the causal effect of vaccination on infection rate? Write down the regression you would run. Explain why.

Question 5 [25 points]

34. Fast Food Sales. Management proposed the following regression model to predict sales at a fast-food outlet.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$$

where

$$x_1 = \text{ number of competitors within one mile}$$
$$x_2 = \text{ population within one mile } (1000\,\text{s})$$
$$x_3 = \{1 \text{ if drive-up window present}$$
$$0 \text{ otherwise}$$
$$y = \text{ sales } (\$1000\,\text{s})$$

The following estimated regression equation was developed after 20 outlets were surveyed.

$$\hat{y} = 10.1 - 4.2x_1 + 6.8x_2 + 15.3x_3$$

a. What is the expected amount of sales attributable to the drive-up window?

b. What are sales for a store with two competitors, a population of 8000 within one mile, and no drive-up window?

c. Suppose I defined x_4 as:

$$\begin{aligned} & x_4=\left\{\begin{array}{l} 1 \text { if drive-up window is \textbf{NOT} present } \\ 0 \text { otherwise } \end{array}\right. \\ \end{aligned}$$

And then I ran the regression:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \epsilon$$

What would happen? Why?

	DF1	α = 0.10																	
DF2	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	Inf
1	39.863	49.5	53.593	55.833	57.24	58.204	58.906	59.439	59.858	60.195	60.705	61.22	61.74	62.002	62.265	62.529	62.794	63.061	63.328
2	8.5263	9	9.1618	9.2434	9.2926	9.3255	9.3491	9.3668	9.3805	9.3916	9.4081	9.4247	9.4413	9.4496	9.4579	9.4662	9.4746	9.4829	9.4912
3	5.5383	5.4624	5.3908	5.3426	5.3092	5.2847	5.2662	5.2517	5.24	5.2304	5.2156	5.2003	5.1845	5.1764	5.1681	5.1597	5.1512	5.1425	5.1337
4	4.5448	4.3246	4.1909	4.1073	4.0506	4.0098	3.979	3.9549	3.9357	3.9199	3.8955	3.8704	3.8443	3.831	3.8174	3.8036	3.7896	3.7753	3.7607
5	4.0604	3.7797	3.6195	3.5202	3.453	3.4045	3.3679	3.3393	3.3163	3.2974	3.2682	3.238	3.2067	3.1905	3.1741	3.1573	3.1402	3.1228	3.105
6	3.776	3.4633	3.2888	3.1808	3.1075	3.0546	3.0145	2.983	2.9577	2.9369	2.9047	2.8712	2.8363	2.8183	2.8	2.7812	2.762	2.7423	2.7222
7	3.5894	3.2574	3.0741	2.9605	2.8833	2.8274	2.7849	2.7516	2.7247	2.7025	2.6681	2.6322	2.5947	2.5753	2.5555	2.5351	2.5142	2.4928	2.4708
8	3.4579	3.1131	2.9238	2.8064	2.7265	2.6683	2.6241	2.5894	2.5612	2.538	2.502	2.4642	2.4246	2.4041	2.383	2.3614	2.3391	2.3162	2.2926
9	3.3603	3.0065	2.8129	2.6927	2.6106	2.5509	2.5053	2.4694	2.4403	2.4163	2.3789	2.3396	2.2983	2.2768	2.2547	2.232	2.2085	2.1843	2.1592
10	3.285	2.9245	2.7277	2.6053	2.5216	2.4606	2.414	2.3772	2.3473	2.3226	2.2841	2.2435	2.2007	2.1784	2.1554	2.1317	2.1072	2.0818	2.0554
11	3.2252	2.8595	2.6602	2.5362	2.4512	2.3891	2.3416	2.304	2.2735	2.2482	2.2087	2.1671	2.1231	2.1	2.0762	2.0516	2.0261	1.9997	1.9721
12	3.1766	2.8068	2.6055	2.4801	2.394	2.331	2.2828	2.2446	2.2135	2.1878	2.1474	2.1049	2.0597	2.036	2.0115	1.9861	1.9597	1.9323	1.9036
13	3.1362	2.7632	2.5603	2.4337	2.3467	2.283	2.2341	2.1954	2.1638	2.1376	2.0966	2.0532	2.007	1.9827	1.9576	1.9315	1.9043	1.8759	1.8462
14	3.1022	2.7265	2.5222	2.3947	2.3069	2.2426	2.1931	2.1539	2.122	2.0954	2.0537	2.0095	1.9625	1.9377	1.9119	1.8852	1.8572	1.828	1.7973
15	3.0732	2.6952	2.4898	2.3614	2.273	2.2081	2.1582	2.1185	2.0862	2.0593	2.0171	1.9722	1.9243	1.899	1.8728	1.8454	1.8168	1.7867	1.7551
16	3.0481	2.6682	2.4618	2.3327	2.2438	2.1783	2.128	2.088	2.0553	2.0282	1.9854	1.9399	1.8913	1.8656	1.8388	1.8108	1.7816	1.7508	1.7182
17	3.0262	2.6446	2.4374	2.3078	2.2183	2.1524	2.1017	2.0613	2.0284	2.0009	1.9577	1.9117	1.8624	1.8362	1.809	1.7805	1.7506	1.7191	1.6856
18	3.007	2.624	2.416	2.2858	2.1958	2.1296	2.0785	2.0379	2.0047	1.977	1.9333	1.8868	1.8369	1.8104	1.7827	1.7537	1.7232	1.691	1.6567
19	2.9899	2.6056	2.397	2.2663	2.176	2.1094	2.058	2.0171	1.9836	1.9557	1.9117	1.8647	1.8142	1.7873	1.7592	1.7298	1.6988	1.6659	1.6308
20	2.9747	2.5893	2.3801	2.2489	2.1582	2.0913	2.0397	1.9985	1.9649	1.9367	1.8924	1.8449	1.7938	1.7667	1.7382	1.7083	1.6768	1.6433	1.6074
21	2.961	2.5746	2.3649	2.2333	2.1423	2.0751	2.0233	1.9819	1.948	1.9197	1.875	1.8272	1.7756	1.7481	1.7193	1.689	1.6569	1.6228	1.5862
22	2.9486	2.5613	2.3512	2.2193	2.1279	2.0605	2.0084	1.9668	1.9327	1.9043	1.8593	1.8111	1.759	1.7312	1.7021	1.6714	1.6389	1.6042	1.5668
23	2.9374	2.5493	2.3387	2.2065	2.1149	2.0472	1.9949	1.9531	1.9189	1.8903	1.845	1.7964	1.7439	1.7159	1.6864	1.6554	1.6224	1.5871	1.549
24	2.9271	2.5383	2.3274	2.1949	2.103	2.0351	1.9826	1.9407	1.9063	1.8775	1.8319	1.7831	1.7302	1.7019	1.6721	1.6407	1.6073	1.5715	1.5327
25	2.9177	2.5283	2.317	2.1842	2.0922	2.0241	1.9714	1.9293	1.8947	1.8658	1.82	1.7708	1.7175	1.689	1.659	1.6272	1.5934	1.557	1.5176
26	2.9091	2.5191	2.3075	2.1745	2.0822	2.0139	1.961	1.9188	1.8841	1.855	1.809	1.7596	1.7059	1.6771	1.6468	1.6147	1.5805	1.5437	1.5036
27	2.9012	2.5106	2.2987	2.1655	2.073	2.0045	1.9515	1.9091	1.8743	1.8451	1.7989	1.7492	1.6951	1.6662	1.6356	1.6032	1.5686	1.5313	1.4906
28	2.8939	2.5028	2.2906	2.1571	2.0645	1.9959	1.9427	1.9001	1.8652	1.8359	1.7895	1.7395	1.6852	1.656	1.6252	1.5925	1.5575	1.5198	1.4784
29	2.887	2.4955	2.2831	2.1494	2.0566	1.9878	1.9345	1.8918	1.8568	1.8274	1.7808	1.7306	1.6759	1.6466	1.6155	1.5825	1.5472	1.509	1.467
30	2.8807	2.4887	2.2761	2.1422	2.0493	1.9803	1.9269	1.8841	1.849	1.8195	1.7727	1.7223	1.6673	1.6377	1.6065	1.5732	1.5376	1.4989	1.4564
40	2.8354	2.4404	2.2261	2.091	1.9968	1.9269	1.8725	1.8289	1.7929	1.7627	1.7146	1.6624	1.6052	1.5741	1.5411	1.5056	1.4672	1.4248	1.3769
60	2.7911	2.3933	2.1774	2.041	1.9457	1.8747	1.8194	1.7748	1.738	1.707	1.6574	1.6034	1.5435	1.5107	1.4755	1.4373	1.3952	1.3476	1.2915
120	2.7478	2.3473	2.13	1.9923	1.8959	1.8238	1.7675	1.722	1.6843	1.6524	1.6012	1.545	1.4821	1.4472	1.4094	1.3676	1.3203	1.2646	1.1926
Inf	2.7055	2.3026	2.0838	1.9449	1.8473	1.7741	1.7167	1.6702	1.6315	1.5987	1.5458	1.4871	1.4206	1.3832	1.3419	1.2951	1.24	1.1686	1