ECON 453 - Econometrics
Fall 2023
Exam 2 Practice Problems

These problems are intended to provide you with some additional examples of the types of questions I may ask on the exam, some of them are based on problems I have given on previous exams and assignments. **Note that these are not intended to cover every possible topic/question that could be included on the exam.** Please review the Exam 2 Study Guide for a more comprehensive list of topics. There are also examples you can use in Problem Set 3 and in-class assignment 6 to 10.

1. A linear probability model is used to examine:
    a. Models in which the dependent variable is categorical
    b. Models in which the dependent variable has been transformed with the natural logarithm
    c. Models in which we suspect non-linear relationships between each of the explanatory variables and the dependent variable
    d. Models in which we suspect a large amount of multicollinearity

2. Which of the following phrases should you say most often when interpreting the coefficients in a linear probability model?
    a. "Percent"
    b. "Percentage points"
    c. "Dollars"
    d. "I have no (*expletive*) idea how to interpret this coefficient"

3. According to our discussion of the instrumental variable methodology, researchers have used quarter-of-birth as an instrument because they argued that it will
    a. Directly affect both the amount of education an individual receives and the wage that individual earns
    b. Directly affect neither the amount of education an individual receives nor the wage that individual earns
    c. Directly affect the wage an individual earns but not directly affect the amount of education that individual receives
    d. Directly affect the amount of education an individual receives but not directly affect the wage that individual earns.

4. A typical difference-in-difference study that involves regression analysis is generally going to include which of the following modeling options?
    a. A linear probability model
    b. An interaction term
    c. A quadratic term
    d. All of the above

5. According to our discussion in class, which of these is the primary explanation for why early econometric studies of unemployment and crime rates across cities often produced confusing results?
    a. Previous economic theories about the relationship between unemployment and crime were incorrect
    b. The specific samples of cities that were studied produced this result. When a different sample of cities was selected, the expected results were found
    c. The models failed to account for city-specific characteristics
    d. All of the above

For questions 6 and 7, suppose we use a dataset containing President Obama's approval ratings each month from January 2009 to December 2016. We estimate a first-differences model where the approval rating is our dependent variable, and the unemployment rate is our only explanatory variable, and come up with the following equation: $\Delta y = -0.8 - 2.6(\Delta x)$

6. In the model above, what does the constant term represent
    a. The predicted Obama approval rating if unemployment is 0
    b. The predicted Obama approval rating if the change in unemployment is 0
    c. The predicted change in the Obama approval rating if unemployment is 0
    d. The predicted change in the Obama approval rating if the change in unemployment is 0

7. Suppose that in October 2016 the unemployment rate was 5 and Obama's approval rating was 45. If the unemployment rate was 4 in November 2016, what does the model predict Obama's approval rating should be in November 2016?
    a. 47.6
    b. 46.8
    c. 42.4
    d. 41.6
    e. Not enough information provided to answer

8. Why might we choose to use a de-trended y-variable in a time-series regression model, as opposed to including a time-trend as an additional explanatory variable?
    a. The de-trended model will produce an R^2 value that more realistically states the explanatory power of our model
    b. The de-trended model is less likely to have an issue with serial correlation in the errors
    c. The de-trended model produces less biased coefficients because it specifically controls for trends in the y-variable
    d. All of the above

9. Suppose we are looking at a time series model predicting y as a function of two explanatory variables. Which of the following equations represents a "lagged dependent variable" specification of this model?
    a. $y_t = \beta_0 + \beta_1 x_{1,t} + \beta_2 x_{2,t} + \beta_3 y_{t-1}$
    b. $y_t = \beta_0 + \beta_1 x_{1,t} + \beta_2 x_{2,t} + \beta_3 t$
    c. $y_t = \beta_0 + \beta_1 x_{1,t-1} + \beta_2 x_{2,t-1} + \beta_3 t$
    d. $(y_t - y_{t-1}) = \beta_0 + \beta_1 (x_{1,t} - x_{1,t-1}) + \beta_2 (x_{2,t} - x_{2,t-1})$

10. Suppose we look at a sample of 25 to 35-year-olds from across the country collected by the American Community Survey. We are going to estimate a linear probability model where our dependent variable is whether or not the person has (their own) children that live in the same house as them. This is a dummy dependent variable equal to 1 if the person has at least one kid (living with them), and 0 if not. This is a large sample. To begin with, our explanatory variables will be the person's age, and dummy variables for whether or not the person is female and whether or not the person has a bachelor's degree.

Source	SS	df	MS
Model	10183.6123	3	3394.53742
Residual	71531.8341	331741	.215625546
Total	81715.4464	331744	.246320797

Number of obs = 331745
F(3,331741) =15742.74
Prob > F = 0.0000
R-squared = 0.1246
Adj R-squared = 0.1246
Root MSE = .46435

kids	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
age	.0432984	.0002576	168.10	0.000	.0427936	.0438033
female	.2079692	.0016229	128.15	0.000	.2047885	.21115
bach	-.1197438	.0017025	-70.33	0.000	-.1230806	-.1164069
_cons	-.9234173	.0078232	-118.04	0.000	-.9387505	-.9080841

a. Interpret the coefficients and discuss the intuitive reason for why we are getting each of these estimated results.
b. Predict the y-variable for a 33-year-old male with a college degree and an obsession with the hit television show *Outer Banks.* Give the practical meaning of this estimate.

Now suppose I include an interaction term of the female and bachelor's degree dummy variables. The results of the new model are:

Source	SS	df	MS
Model	10599.332	4	2649.83299
Residual	71116.1144	331740	.214373046
Total	81715.4464	331744	.246320797

Number of obs = 331745
F(4,331740) =12360.85
Prob > F = 0.0000
R-squared = 0.1297
Adj R-squared = 0.1297
Root MSE = .463

kids	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
age	.0432021	.0002568	168.21	0.000	.0426987	.0437055
female	.2601139	.0020051	129.72	0.000	.256184	.2640439
bach	-.0415728	.0024562	-16.93	0.000	-.0463868	-.0367588
fem_bach	-.1495287	.0033955	-44.04	0.000	-.1561839	-.1428736
_cons	-.9438447	.0078142	-120.79	0.000	-.9591603	-.9285291

c. Describe the practical meaning of the results from the interaction term. Then present separate prediction equations for males and females.
d. Using this second model, I predicted the y-hat value for each of the 331,745 people in the sample (by hand). I then created a variable "prob_kids" which I set equal to 0 if the person's y-hat value was less than 0.5, and equal to one if y-hat was greater than (or equal to) 0.5. Use the two-way table below to discuss how accurate our model is. How good are we at predicting for those that actually do not have kids? How good are we at predicting for those that actually do have kids? How good are we at predicting overall?

	prob_kids		
kids	0	1	Total
0	140,417	45,580	185,997
1	69,068	76,680	145,748
Total	209,485	122,260	331,745

11. Suppose now we restrict the sample in problem 2 to 25 to 35-year-olds living in Idaho. We are going to estimate a logistic regression (logit) model where our dependent variable is whether or not the person has (their own) children that live in the same house as them. This is a dummy dependent variable equal to 1 if the person has at least one kid (living with them), and 0 if not. Our explanatory variables will be the person's age, income (in thousands of dollars), and dummy variables for whether or not the person is female and whether or not the person has a bachelor's degree.

```
File  Edit  Tests  Save  Graphs  Analysis  LaTeX

Model 3: Logit, using observations 100321-101899 (n = 1579)
Dependent variable: kids
Standard errors based on Hessian

             coefficient   std. error      z         slope
  -----------------------------------------------------------
  const      -5.64031      0.548130     -10.29
  age         0.185272     0.0182460     10.15      0.0447886
  income      0.00401708   0.00245257     1.638     0.00097111(
  female      0.850850     0.116150       7.325     0.202225
  bach       -0.283333     0.125085      -2.265    -0.0691169
```

a. Interpret the coefficients and discuss the intuitive reason for why we are getting each of these estimated results. Do the results match your expectations?

b. Suppose you run into a gentleman who goes by the name T-bone. T-bone is a 32-year-old proud Idaho native who works in construction and likes to paint pictures of Idaho's breathtaking wildflowers. He once tried crystal meth in college (while earning his bachelor's degree in art), but did not care for it. If T-bone brings in a total of $45,000 in annual income, what is the predicted probability that he has children? What happens to this predicted probability if he had not earned his bachelor's degree?

12. Consider a 2017 study in the Journal of the American Medical Association that examines whether legalization of recreational marijuana in Colorado and Washington affected adolescent usage.

a. The study used a large survey of 8^{th} graders that is conducted on an annual basis. They looked at the averages for each state in the few years before legalization and the few years after. The main question is: did you use marijuana in the past month? The numbers in the table below show the percentage that said "Yes" to this question. Using the information in the table, what is the difference-in-difference estimate of the impact of legalization on 8^{th} grade usage rates in Colorado? In Washington? Do these findings make sense?

State	Before Legalization	After Legalization
Colorado	8.9	8.9
Washington	6.2	8.2
All other states	7.6	6.3

b. Discuss how comfortable you are with the control group being "all other states". Is there a better group to use, or is this the best option?

c. The "before" period of the study is 2010-2012 and the "after" period is 2013-2015. Explain why it might be useful to examine survey data from earlier surveys, and how this might influence the analysis.

13. Consider a time-series dataset that collects information on U.S. average life expectancy over time. The data is collected annually for 1961-2011. The y-variable in our analysis will be the average life expectancy in years. The explanatory variables we will use are: the % of the population that has a bachelor's degree (0 to 100 scale), GDP per capita (in thousands of $), and average cigarette consumption per capita (in thousands of cigarettes per year). The results of estimating a simple static model are presented below:

```
Model 1: OLS, using observations 1961-2011 (T = 51)
Dependent variable: LifeExp

               coefficient   std. error   t-ratio    p-value
  ----------------------------------------------------------
  const         63.8859      1.01600      62.88     5.13e-047 ***
  Bachelors      0.621095    0.0336735    18.44     3.70e-023 ***
  GDPpc1000s    -0.108092    0.0294798    -3.667    0.0006    ***
  Cigs           0.578981    0.157551      3.675    0.0006    ***

Mean dependent var   74.27262   S.D. dependent var   2.696357
Sum squared resid    2.433622   S.E. of regression   0.227550
R-squared            0.993305   Adjusted R-squared   0.992878
F(3, 47)             2324.507   P-value(F)           4.45e-51
Log-likelihood       5.216481   Akaike criterion    -2.432961
Schwarz criterion    5.294341   Hannan-Quinn         0.519869
rho                  0.225086   Durbin-Watson        1.450097
```

a. Interpret the coefficients on each of the explanatory variables. Comment on the results of this model. How do you feel about the validity of the results overall, and do they match your expectations?

b. Using the results below, conduct a standard test for first-order serial correlation in the model from part a. What do these results tell us?

gretl: Durbin-Watson

```
Durbin-Watson statistic = 1.4501

H1: positive autocorrelation
   p-value = 0.0071632
H1: negative autocorrelation
   p-value = 0.992837
```

c. Being that this is a time-series dataset, I next decide to include a time trend variable. This is a variable that starts at t=1 in 1961, t=2 in 1962, etc. The results of this modeling specification are presented below. Interpret the meaning of the estimated coefficient on the trend variable. Comment on how the results have changed as compared to the static model. Do you feel that this was an improvement to our model?

```
Model 2: OLS, using observations 1961-2011 (T = 51)
Dependent variable: LifeExp

              coefficient   std. error   t-ratio    p-value
  ---------------------------------------------------------
  const        64.5491      1.12597      57.33     1.95e-044 ***
  Bachelors     0.533572    0.0741604     7.195    4.65e-09  ***
  GDPpc1000s   -0.127445    0.0327076    -3.896    0.0003    ***
  Cigs          0.630883    0.161169      3.914    0.0003    ***
  time          0.0556094   0.0420669     1.322    0.1927

Mean dependent var   74.27262   S.D. dependent var   2.696357
Sum squared resid    2.344555   S.E. of regression   0.225762
R-squared            0.993550   Adjusted R-squared   0.992990
F(4, 46)             1771.544   P-value(F)           9.93e-50
Log-likelihood       6.167251   Akaike criterion    -2.334502
Schwarz criterion    7.324626   Hannan-Quinn         1.356537
rho                  0.226787   Durbin-Watson        1.405284
```

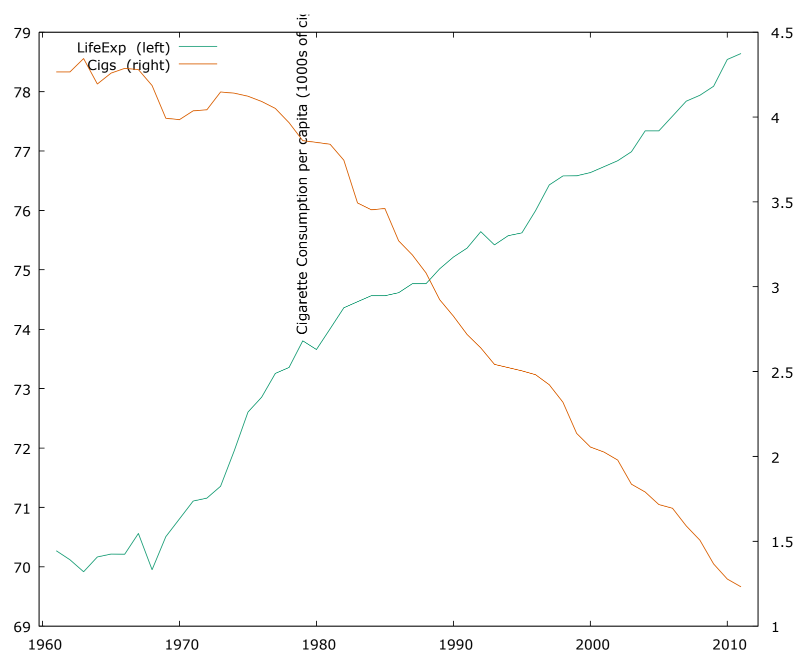
d. Suppose it is a typical day, and Dan has a brilliant idea. “We should be using lagged values of life expectancy as an explanatory variable” he thinks to himself. Furthermore, why not try each of the three explanatory variables in separate models, instead of all in the same model? Dan is on a roll, so he runs the regressions and obtains the results below. Discuss the idea of this type of modeling, and what these particular results tell us.

```
Model 3: OLS, using observations 1962-2011 (T = 50)
Dependent variable: LifeExp

              coefficient   std. error   t-ratio    p-value
  ---------------------------------------------------------
  const        13.6741      6.56139      2.084     0.0426    **
  Bachelors     0.0813211   0.0389405    2.088     0.0422    **
  LifeExp_1     0.796686    0.0985381    8.085     1.90e-010 ***

Mean dependent var   74.35265   S.D. dependent var   2.661830
Sum squared resid    2.241281   S.E. of regression   0.218373
R-squared            0.993544   Adjusted R-squared   0.993270
F(2, 47)             3616.724   P-value(F)           3.42e-52
Log-likelihood       6.677454   Akaike criterion    -7.354909
Schwarz criterion   -1.618840   Hannan-Quinn        -5.170581
rho                  0.010252   Durbin's h           0.101064
```

```
Model 4: OLS, using observations 1962-2011 (T = 50)
Dependent variable: LifeExp

              coefficient   std. error   t-ratio    p-value
  ----------------------------------------------------------
  const        3.05035      3.33725       0.9140   0.3654
  GDPpc1000s   0.0137216    0.0149071     0.9205   0.3620
  LifeExp_1    0.955419     0.0509919    18.74     1.93e-023 ***

Mean dependent var   74.35265   S.D. dependent var   2.661830
Sum squared resid    2.405881   S.E. of regression   0.226250
R-squared            0.993070   Adjusted R-squared   0.992775
F(2, 47)             3367.677   P-value(F)           1.81e-51
Log-likelihood       4.905750   Akaike criterion    -3.811499
Schwarz criterion    1.924570   Hannan-Quinn        -1.627171
rho                 -0.061577   Durbin's h          -0.466817
```

```
Model 5: OLS, using observations 1962-2011 (T = 50)
Dependent variable: LifeExp

             coefficient   std. error   t-ratio    p-value
  ---------------------------------------------------------
  const      -0.441961     3.16751      -0.1395   0.8896
  Cigs        0.0178689    0.101173      0.1766   0.8606
  LifeExp_1   1.00747      0.0386977    26.03     1.43e-029 ***

Mean dependent var   74.35265   S.D. dependent var   2.661830
Sum squared resid    2.447627   S.E. of regression   0.228204
R-squared            0.992950   Adjusted R-squared   0.992650
F(2, 47)             3309.837   P-value(F)           2.71e-51
Log-likelihood       4.475670   Akaike criterion    -2.951339
Schwarz criterion    2.784730   Hannan-Quinn        -0.767012
rho                 -0.098900   Durbin's h          -0.727080
```

e. Suppose that our whole goal in constructing this model was to determine how the rate of cigarette use in the U.S. affects the overall health of the U.S. population over time. Consider the time plot below and discuss whether we can determine this using time-series data on life expectancy and cigarette use. What other methods/samples/models/variables could we try to determine the link between rates of cigarette consumption and health?



14. Consider a dataset that looks at 269 universities in the U.S. This is a set of research schools that offer bachelor's, master's, *and* doctoral degrees. We collect information from 2003 to 2011 (9 years of observations). Suppose we want to look at the relationship between enrollment and tuition. Our dependent variable will be the undergraduate enrollment in the school (in thousands of students), and our explanatory variables will be average undergraduate tuition (in thousands of dollars) and a dummy equal to 1 if the school is public and 0 if it is private.

    a. Suppose we decide to start with a simple cross-sectional model of the 269 schools in the most recent year in our data, 2011. Interpret the coefficients and discuss whether or not they match your expectations.

```
Model 1: OLS, using observations 1-269
Dependent variable: enroll

             coefficient   std. error   t-ratio    p-value
  ---------------------------------------------------------
  const      -0.842269     2.82025      -0.2987   0.7654
  tuition2    0.244865     0.101444      2.414    0.0165    **
  public     18.4178       2.36624       7.784    1.57e-013 ***

Mean dependent var   12.66747   S.D. dependent var   9.994487
Sum squared resid    14847.69   S.E. of regression   7.471169
R-squared            0.445370   Adjusted R-squared   0.441200
F(2, 266)            106.7997   P-value(F)           8.96e-35
Log-likelihood      -921.1589   Akaike criterion     1848.318
Schwarz criterion    1859.102   Hannan-Quinn         1852.649
```

    b. Next, we run a pooled model where we include information for all schools for all 9 years. Consider the results of this model. Interpret the coefficients and compare the results to those in part a. Has panel data improved our analysis?

```
Model 4: Pooled OLS, using 2421 observations
Included 269 cross-sectional units
Time-series length = 9
Dependent variable: enroll

             coefficient   std. error   t-ratio    p-value
  ---------------------------------------------------------
  const      16.8570       0.237417      71.00     0.0000    ***
  tuition2    0.170388     0.0316499      5.384    8.01e-08  ***
  public    -15.4649       0.698584     -22.14     4.97e-099 ***

Mean dependent var   11.83702   S.D. dependent var   9.292279
Sum squared resid    120052.2   S.E. of regression   7.046232
R-squared            0.425473   Adjusted R-squared   0.424998
F(2, 2418)           895.3401   P-value(F)           1.0e-291
Log-likelihood      -8160.735   Akaike criterion     16327.47
Schwarz criterion    16344.85   Hannan-Quinn         16333.79
```

All of a sudden it hits us: we have panel data, we should probably run a fixed effects model! We decide to now run a model where we drop the public dummy variable, but add school fixed effects and dummies for each year. The results are displayed below. Note: the dummies for time omit the first year (2003). “dt_2” is the dummy for year 2 (2004) and so on.

Model 3: Fixed-effects, using 2421 observations
Included 269 cross-sectional units
Time-series length = 9
Dependent variable: enroll

	coefficient	std. error	t-ratio	p-value	
const	12.9586	0.211591	61.24	0.0000	***
tuition2	−0.146817	0.0170837	−8.594	1.59e-017	***
dt_2	0.122868	0.0791117	1.553	0.1205	
dt_3	0.313031	0.0798440	3.921	9.11e-05	***
dt_4	0.542660	0.0814041	6.666	3.33e-011	***
dt_5	0.766293	0.0826279	9.274	4.24e-020	***
dt_6	1.12916	0.0879677	12.84	2.18e-036	***
dt_7	1.44669	0.0932414	15.52	1.55e-051	***
dt_8	1.70160	0.0962342	17.68	2.04e-065	***
dt_9	2.02079	0.102105	19.79	3.35e-080	***

Mean dependent var	11.83702	S.D. dependent var	9.292279
Sum squared resid	1791.109	S.E. of regression	0.914218
LSDV R-squared	0.991428	Within R-squared	0.267333
LSDV F(277, 2143)	894.8322	P-value(F)	0.000000
Log-likelihood	−3070.471	Akaike criterion	6696.942
Schwarz criterion	8307.101	Hannan-Quinn	7282.460
rho	0.777168	Durbin-Watson	0.431830

c. Interpret the coefficient on the tuition variable in this new model. Explain whether this result makes more sense, and why the change occurred.

d. Discuss what the time dummy variables tell us in this model.