Total points: 100

Due Date: Wednesday, April 22 by 11:59 p.m. EST

*Directions:* This final exam requires you to use the statistical software R. To make grading exams easier, I require that you complete the exam using R-markdown denoting each question with “# Question 1” followed by each part with “# Part A” etc. and include the code for each part in one code chunk. You will turn in a pdf file (i.e. knit the file as a Word document and then save as a pdf file) and the R-markdown file on Canvas by the due date and time. You are to work **independently** on this exam; however, you are welcome to use anything available on Canvas. If I find evidence that you worked with others, you will receive a zero for this exam. I will NOT accept late exams, so please plan accordingly. If you have any questions please do not hesitate to ask me.

1. The U.S. Department of Health and Human Services (HHS) believes that increasing taxes on cigarettes would be an effective policy tool to decrease cigarette consumption, and therefore reduce illness and death from smoking. However, HHS wants to know precisely how much the price of cigarettes must increase to induce a substantial reduction in cigarette consumption. For example, by how much would the price of cigarettes need to increase to reduce cigarette consumption by 25%? To answer this question, HHS hires you as a consultant to obtain a good estimate of the price elasticity of demand for cigarettes.

# Data

The data you will use for your study are contained in the data file “cigarettes” located on Canvas in the “Final Exam” folder. This panel dataset consists of annual data for each of the 48 contiguous U.S. states for the years 1985 to 1995. *cigs* is annual per capita cigarette sales in packs; *price* is the average price of a pack of cigarettes in cents; *inc* is state personal income in dollars; *taxe* is the federal, state, and local excise tax per pack of cigarettes in cents; *taxs* is the sales tax per pack of cigarettes in cents; *state* is state; *sindex* is a state index; *year* is year. All prices, income, and taxes are adjusted for inflation using the consumer price index, and therefore are in real terms.

Given the data available to you, to analyze the relationship between the price of cigarettes and cigarette consumption, you specify the following demand and supply equations for cigarettes.

Demand Equation: *lcigsit* = *β*0 + *β*1*lpriceit* + *β*2*lincit* + *u*1*it* (1)

Supply Equation: *lcigsit* = *α*0 + *α*1*lpriceit* + *α*2*taxeit* + *α*3*taxsit* + *u*2*it* (2)

where the letter “l” before a variable designates natural logarithm, and the subscripts i and t designate state and year. Note that *inc*, *cigs*, and *price* are in logarithmic form and *taxe* and *taxs* are not in logarithmic form.

1. (*5 points*) For each of the following variables, create a new variable which is the logarithm of the variable: *inc*, *cigs*, *price*. Designate each of these variables by the first letter “l”, that is *linc*, *lcigs*, *lprice*. (No written answer required).
2. (*5 points*) What coefficient measures the price elasticity of demand for cigarettes? If the objective of your study is to estimate the price elasticity of demand for cigarettes, then why does your model have a cigarette supply equation? Carefully explain.
3. (*5 points*) Does the demand equation satisfy the order condition, in particular, is the demand equation exactly identified, underidentified or overidentified? Explain how you determined this? Do you believe the exclusion restrictions for the demand equation are valid? That is, do you believe the exogenous variable(s) that have been excluded from the demand equation to identify it can be validly excluded? Yes/no. Explain.
4. (*5 points*) Estimate the demand equation (1) using the pooled OLS estimator and cluster robust standard errors (use small sample correction in all cluster robust standard errors for this exam). Report the results.
5. (*5 points*) Estimate the demand equation (1) using the 2SLS estimator and cluster robust standard errors. Report the results. What variable(s) are identifying instruments for the 2SLS estimator? Do you believe the identifying instruments are relevant and exogenous (no tests are necessary just explain your reasoning)? Yes or no. Explain.
6. (*5 points*) Compare the OLS and 2SLS estimates of *β*1. Does the OLS estimate appear to be biased up or down relative to the 2SLS estimate? Does the bias appear to be relatively large or small?
7. (*5 points*) Estimate the first-stage regression for *lprice* using the pooled OLS estimator and cluster robust standard errors. Do the estimates of the coefficients of the identifying instruments have expected signs? Yes or no. Explain. Check for instrument relevance. Do you believe the instruments are relatively weak or strong?
8. (*5 points*) Now test the hypothesis that *lprice* is exogenous in demand equation (1) using the endogeneity test. Interpret the result. That is, what does the test results suggest about the exogeneity of *lprice*? What does this tell you about the bias and consistency in the OLS estimate of *β*1?
9. (*5 points*) Test the overidentifying restrictions for the demand equation (1). Interpret the result. What does this tell you about the validity of your instruments? Does the result of your test provide evidence that the instruments are exogenous or endogenous?
10. (*5 points*) Because you have panel data, you decide to estimate a demand equation with fixed-effects and time effects. You specify the following model.

*lcigsit* = *β*0 + *β*1*lpriceit* + *lincit* + *ai* + *θt* + *u*1*it* (3)

Estimate this demand equation using the fixed effects 2SLS estimator with cluster robust standard errors. Report the results for *lprice* and *linc* (Do not report the estimates of the coefficients of the state and year dummy variables). Compare the estimate of *β*1 for the 2SLS model with fixed and time effects and the 2SLS model without fixed and time effects that you reported previously. Are the estimates similar or are they noticeably different?

1. (*5 points*) Based on your results in part j, describe the general trend in smoking over time. Explain how you know this.
2. (*5 points*) You have estimated the following models: OLS, 2SLS,and 2SLS with fixed and time effects. Choose the model that you believe gives you the best estimate of the price elasticity of demand for cigarettes. Carefully explain why you chose this model. Do you believe that this estimate of the price elasticity of demand for cigarettes is a relatively good or relatively poor estimate? Explain.
3. (*5 points*) The average price of a pack of cigarettes in the U.S. is $6.00. Use the estimate of the price elasticity of demand from the model you selected in part ?? to answer the following policy question. By how much would the average price of cigarettes have to increase to reduce cigarette consumption by 25%? Explain how you obtained your answer.

2. You are employed as an economist at the American Diabetes Association (ADA). The ADA wants to better understand how body weight, lifestyle choices, and socioeconomic factors affect the likelihood of getting diabetes. It also wants to predict the likelihood that an individual with particular characteristics will be diabetic.

# Data

The data for your study are contained in the dataset “diabetes” located on Canvas in the “Final Exam” folder. It is a random sample of 5,051 adults in the U.S. between age 20 and 80. The variables are as follows. *male* is a dummy variable that takes a value of 1 if male and 0 otherwise. *age* is age measured in years. *marr* is a dummy variable that takes a value of 1 if married and 0 otherwise. *inc* is family income measured in thousands of dollars. *coll* is a dummy variable that takes a value 1 if a college education and 0 otherwise. *alc* is a dummy variable that takes a value of 1 if consumption of alcohol and 0 otherwise. *cig* is a dummy variable that takes a value of 1 if smoke cigarettes and 0 otherwise. *exer* is a dummy variable that takes a value of 1 if recreational exercise and 0 otherwise. *owgt* is a dummy variable that takes a value of 1 if overweight defined as a body mass index of 25 to less than 30 and 0 otherwise. *obese* is a dummy variable that takes a value of 1 if obese defined as a body mass index of 30 or greater and 0 otherwise. *diab* is a dummy variable that takes a value of 1 if diabetes and 0 otherwise.

The dependent variable is *diab*. You choose the following explanatory variables to include in the model. Body weight is measured by the variables *owgt* and *obes*e. Lifestyle choice variables are *exer*, *cig*, and *alc*. Socioeconomic variables are *inc*, *coll*, and *marr*. You also include the personal characteristic variables *male* and *age*. There are two reasons why you include these two variables. (1) You believe they may be potential confounding variables. (2) You believe they will allow you to make better predictions of the likelihood of developing diabetes.

You decide to estimate three alternative statistical models: (1) Linear probability model using the OLS estimator and White robust standard errors. (2) Logit model using the maximum likelihood estimator. (3) Probit model using the maximum likelihood estimator. Given your knowledge of these models and the results, you will choose what you believe is the most appropriate model to draw conclusions and make predictions. Your primary interest is in estimated the following response probability:

*P*(*diab* = 1|*x*) (4)

1. (*5 points*) Estimate three models. (1) Linear probability model using the OLS estimator and White robust standard errors. (2) Logit model using the maximum likelihood estimator. (3) Probit model using the maximum likelihood estimator. Report the results for each model. No writing required.
2. (*5 points*) Calculate the Average partial effects (APE) for the logit and probit estimates. Compare these results with those of the OLS LPM. Are there any noticeable differences?
3. (*5 points*) To measure how well each model fits the sample data, calculate the Macfadden’s Pseudo *R*2. Is the goodness of fit of the three models similar or noticeably different? Which model fits the data best? How well do you think the three models fit the sample data? Explain.
4. (*5 points*) As an alternate measure of goodness-of-fit calculate the percentage of correct predictions for the probit and logit model results. Is the goodness of fit of the two models similar or noticeably different? Which model fits the data best? How well do you think the two models fit the sample data? Explain.
5. (*5 points*) Choose the model that you believe is the best model to make conclusions and predictions. Use the results from the model you selected to make conclusions about the effects of body weight, lifestyle choices, and socioeconomic characteristics on diabetes (i.e. sign and significance).
6. (*5 points*) Use the model you selected to predict the probability that an obese, age 50, married male, with $50 (thousand) annual income, who doesn’t exercise, drink alcohol, or smoke cigarettes, with less than a college education will be diabetic.
7. (*5 points*) Repeat the calculation in part g only assume the person is not obese.

What is the predicted probability in this case?