Notes: The analyses for this test are to be performed using Stata. In answering all questions, note what commands you used (if any) including the variables included and any options specified. Write your answers to each question in the space provided. If necessary, use space on the back of the page, clearly identifying which question is applicable and noting under the question itself that the answer continues elsewhere. You may use Stata help files, but you may not use any internet help for the Stata commands. You may use your text, class notes, problem sets (including my logs and do files if you have previously printed them out), but you may not use any internet resources. You may only use the internet to access the datasets as indicated below.

There are two sections to this exam. Section one, containing questions 1-10, is potentially worth 50 points in total, and each question is potentially worth 5 points. Section two, containing questions 11-15, is also potentially worth 50 points in total, and each question is potentially worth 10 points.
For the questions in this section, use the dataset available at [http://www.law.upenn.edu/fac/jklick/test1a.txt](http://www.law.upenn.edu/fac/jklick/test1a.txt) which contains the following variables:

- **diabetes**: an indicator taking the value of 1 if the individual has diabetes
- **income**: a categorical measure of income (1 is lowest, 8 is highest)
- **age**: individual’s age measured in years
- **female**: an indicator taking the value of 1 if the individual is female
- **preg**: an indicator taking the value of 1 if the individual is pregnant
- **educa**: a categorical measure of education (1 is lowest, 6 is highest)
- **smoke**: an indicator taking the value of 1 if the individual smokes
- **married**: an indicator taking the value of 1 if the individual is married
- **sepdiv**: an indicator taking the value of 1 if the individual is separated or divorced
- **children**: the number of children the individual has
- **unemployed**: an indicator taking the value of 1 if the individual is unemployed
- **bmi**: body mass index

1. Run a regression of bmi on all available variables (including age and income as quadratics) except diabetes. Is there evidence of heteroskedasticity in the relationship? How do you know?

2. Should you do something to correct for any heteroskedasticity in the relationship estimated above? How should you correct for it and why? Incorporate your answer for this question when answering subsequent questions in this section.
3. What is the marginal effect of age on bmi? At what type 1 error level is this relationship statistically significant?

4. Re-run your regression allowing for differential effects of being married by sex. Is the effect of being married on bmi (statistically) significantly different for men and women? How do you know?

5. Assume that bmi is measured with an additive random error, where the error is mean zero and variance > 1. Relative to the results you estimated for the questions above, what differences would you expect to find if you were able to estimate the same regressions using true (i.e., bmi measured without error) as your dependent variable? Why? Would it make a difference if the measurement error were not mean zero? Why?
6. Assume that bmi is measured with an additive random error, where the error is mean zero and variance > 1. However, for a 10% random subsample of the data, you have true bmi (e.g., the survey organization sent out monitors to measure height and weight precisely). Should you use this information when estimating your regression? If so, how and why?

7. In a linear probability regression of diabetes on all of the available variables, consider the following 2 cases: 1) bmi is measured with a random mean zero variance 25 error; 2) age is measured with a random mean zero variance 25 error. What problem does the measurement error generate in each case? In which case should that problem be more severe and why?

8. Estimate a regression of diabetes status on all of the variables in the original dataset using OLS, logit, and probit. Which model implies the largest (in magnitude) effect of an increase in age by one year for a person with the mean characteristics in the sample? Are there any reasons to prefer one model to the others?
9. If you wanted to model the number of children an individual has using the income, education, and married variables, what empirical model(s) would be appropriate? Provide justifications for your decisions, run the model(s), and interpret the results.

10. Pick one of the questions in this section (1-9) to count double.
For the questions in this section, use the dataset available at http://www.law.upenn.edu/fac/jklick/test1b.txt which contains the following variables:

date: The calendar date
prc: The price at which the last trade of the day was transacted
sprtrn: The return on the S&P 500 for the day
adjprc: prc adjusted for stock splits and dividends
rf: The risk free rate of return

The dataset contains share price information for Dura Pharmaceuticals (ticker DURA). It is alleged that Dura made misleading statements about its earnings and the likelihood of securing regulatory approval for a new asthma drug on April 15, 1997. It is further alleged that when Dura disclosed the truth about earnings as well as information that suggested the asthma drug would not, in fact, be approved, stock price suffered. This disclosure was made after trading had ended on February 24, 1998.

11. Securities fraud cases require an element known as “transaction causation” which essentially requires a showing that the alleged wrong-doing (i.e., the misleading statements) causally affected share price. In a Supreme Court case examining Dura, the Court also held that “loss causation” must be demonstrated by showing that when the market was informed that the earlier statement was not true, market price again reacted. In both cases, causation requires that the statement/disclosure led to any price change, not normal market or firm volatility. Provide an analysis that examines both transaction causation and loss causation. Use only the 100 trading days before the misleading statement to estimate your counterfactual model for both transaction causation and loss causation (i.e., use the same pre-event days for both events).
12. Is it possible to parse out the effects of the earnings claims and the regulatory approval statements? That is, suppose DURA’s defense can credibly argue that the earnings misstatements were simple mistakes (and therefore not fraudulent statements giving rise to a legal claim) but there is no such defense for the regulatory approval claim (i.e., there is strong evidence that DURA knew the approval claims were false at the time they were made and therefore fraudulent). If the causal analyses require that the stock price movements be tied to the specific fraud, as opposed to other plausible but non-prohibited causal mechanisms, can regression analysis meet this standard given the facts as stated above? Why or why not? Now suppose the disclosure on February 24, 1998 concerned only the fraudulent statement about FDA approval but not the earnings misstatements (i.e., assume the misstatement was corrected before then). How might regression analysis be used by the defense to argue that the event studies actually cut against an inference of loss causation? Would the shareholders have any argument against this defense?

13. Re-do the analysis from Question 11, but use the method suggested by Gelbach, Helland, and Klick (from the article on your syllabus) to perform the statistical inference. Why might this approach be better than the standard event study approach?
14. The capital asset pricing model suggests that the return on an asset $i$ is described as:

$$ r_{it} = r_{ft} + \beta \left( r_{mt} - r_{ft} \right) + \varepsilon_{it} $$

where $r_{mt}$ is the return for the market portfolio for the given day and $r_{ft}$ is the risk free return. Use the capital asset pricing model to do the analyses required for Question 11.

15. Choose a question from this section (11-14) to count double.