Midterm 2

Due: Wednesday, July 9 by 11:59 pm.

Please submit as a single pdf document on Canvas.

All the data sets are on Canvas.

1. [25 points] Use the data in **FERTIL2.XLSX** to answer this question.
   a. Estimate the model
   $$children = \beta_0 + \beta_1 age + \beta_2 age^2 + \beta_3 educ + \beta_4 electricity + \beta_5 urban + u$$
   And report the usual and heteroskedasticity-robust standard errors. Are the robust standard errors always bigger than the nonrobust ones?
   b. Add the three religious dummy variables and test whether they are jointly significant. What are the p-values for the nonrobust and robust tests?
   c. From the regression in part (b), obtain the fitted values $\hat{y}$ and the residuals, $\hat{u}$. Regress $\hat{u}^2$ on $\hat{y}$, $\hat{y}^2$ and test the joint significance of the two regressors.

2. [25 points] Use the data set **Movies** on Canvas.
   Does viewing a violent movie lead to violent behavior? If so, the incidence of violent crimes, such as assaults, should rise following the release of a violent movie that attracts many viewers. Alternatively, movie viewing may substitute for other activities (such as alcohol consumption) that lead to violent behavior, so that assaults should fall when more viewers are attracted to the cinema. The dataset includes weekend U.S. attendance for strongly violent movies (such as Hannibal), mildly violent movies (such as Spider-Man), and nonviolent movies (such as Finding Nemo). The dataset also includes a count of the number of assaults for the same weekend in a subset of counties in the United States. Finally, the dataset includes indicators for year, month, whether the weekend is a holiday, and various measures of the weather.
   a.
      i. Regress the logarithm of the number of assaults [ln_assaults = ln(assaults)] on the year and month indicators. Is there evidence of seasonality in assaults? That is, do there tend to be more assaults in some months than others? Explain.
      ii. Regress total movie attendance (attend = attend_v + attend_m + attend_n) on the year and month indicators. Is there evidence of seasonality in movie attendance? Explain.
   b. Regress ln_assaults on attend_v, attend_m, attend_n, the year and month indicators, and the weather and holiday control variables available in the data set.
      i. Based on the regression, does viewing a strongly violent movie increase or decrease assaults? By how much? Is the estimated effect statistically significant?
      ii. Does attendance at strongly violent movies affect assaults differently than attendance at moderately violent movies? Differently than attendance at nonviolent movies?
      iii. A strongly violent blockbuster movie is released, and the weekend's attendance at strongly violent movies increases by 6 million; meanwhile, attendance falls by 2 million for moderately violent movies and by 1 million for nonviolent movies. What is the predicted effect on assaults? Construct a 95% confidence interval for the change in assaults.

c. It is difficult to control for all the variables that affect assaults and that might be correlated with movie attendance. For example, the effect of the weather on assaults and movie attendance is only crudely approximated by the weather variables in the data set. However, the data set does include a set of instruments, pr_attend_v, pr_attend_m, and pr_attend_n, that are correlated with attendance but are (arguably) uncorrelated with weekend-specific factors (such as the weather) that affect both assaults and movie attendance. These instruments use historical attendance patterns, not information on a particular weekend, to predict a film's attendance in a given weekend. For example, if a film's attendance is high in the second week of its release, then this can be used to predict that its attendance was also high in the first week of its release. (The details of the construction of these instruments are available in the Dahl and DellaVigna's paper on Canvas) Run the regression from part (b) (including year, month, holiday, and weather controls) but now using pr_attend_v, pr_attend_m, and pr_attend_n as instruments for attend_v, attend_m, and attend_n. Use this regression to answer (b)(i)–(b)(iii).

d. The intuition underlying the instruments in (c) is that attendance in a given week is correlated with attendance in surrounding weeks. For each move category, the data set includes attendance in surrounding weeks. Run the regression using the instruments attend_v_f, attend_m_f, attend_n_f, attend_v_b, attend_m_b, and attend_n_b instead of the instruments used in part (c). Use this regression to answer (b)(i)–(b)(iii).

e. Based on your analysis, what do you conclude about the effect of violent movies on (short-run) violent behavior?

3. [25 points] We examined **Koop and Tobias's data** on wages, education, ability, and so on. We considered the model

$$\begin{aligned}\ln \mathit{Wage} = \beta_1 &+ \beta_2\, \mathit{Educ} + \beta_3\, \mathit{Ability} + \beta_4\, \mathit{Experience} \\ &+ \beta_5\, \mathit{Mother's\ education} + \beta_6\, \mathit{Father's\ education} + \beta_7\, \mathit{Broken\ home} \\ &+ \beta_8\, \mathit{Siblings} + \varepsilon.\end{aligned}$$

a. We are interested in possible nonlinearities in the effect of education on ln Wage. (Koop and Tobias focused on experience. As before, we are not attempting to replicate their results.) A histogram of the education variable shows values from 9 to 20, a spike at 12 years (high school graduation), and a second at 15. Consider aggregating the education variable into a set of dummy variables:

HS = 1 if Educ $\leq$ 12, 0 otherwise
Col = 1 if Educ > 12 and Educ $\leq$ 16, 0 otherwise
Grad = 1 if Educ >16, 0 otherwise.

Replace Educ in the model with (Col, Grad), making high school (HS) the base category, and recompute the model. Report all results. How do the results change? Based on your results, what is the marginal value of a college degree? What is the marginal impact on ln Wage of a graduate degree?

b. The aggregation in part (a) actually loses quite a bit of information. Another way to introduce nonlinearity in education is through the function itself. Add $Educ^2$ to the equation in part (a) and recompute the model. Again, report all results. What changes are suggested? Test the hypothesis that the quadratic term in the equation is not needed—that is, that its coefficient is zero. Based on your results, sketch a profile of log wages as a function of education.

c. One might suspect that the value of education is enhanced by greater ability. We could examine this effect by introducing an interaction of the two variables in the equation. Add the variable

$$EducAb = Educ \times Ability$$

to the base model in part a. Now, what is the marginal value of an additional year of education? The sample mean value of ability is 0.052374. Compute a confidence interval for the marginal impact on ln Wage of an additional year of education for a person of average ability.

d. Combine the models in (b) and (c). Add both $Educ^2$ and $EducAb$ to the base model in the beginning of the question and reestimate. As before, report all results and describe your findings. If we define low ability as less than the mean and high ability as greater than the mean, the sample averages are -0.798563 for the 7,864 low-ability individuals in the sample and +0.717891 for the 10,055 high-ability individuals in the sample. Using the formulation in part (b), with this new functional form, sketch, describe, and compare the log wage profiles for low- and high-ability individuals.

e. Suppose that you are now given the following regression model:

$$\begin{aligned}\ln Wage = \beta_1 + \beta_2 Educ \times I(Educ < \tau) + \beta_3 Educ \times I(Educ \geq \tau) + \beta_4 Exper \\ + \beta_5 Mother'sEducation + \beta_6\, Father'sEducation + \beta_7 Brokenhome \\ + \beta_8 Siblings + \varepsilon\end{aligned}$$

where τ is the threshold parameter, $I(Educ < \tau) = 1$ if $Educ < \tau$, and zero otherwise. Similarly, $I(Educ \geq \tau) = 1$ if $Educ \geq \tau$, and zero otherwise. This is known as a threshold regression and education is the threshold variable. Estimate the threshold and the model parameters. Show your work and estimation output.

4. [25 points] Using the **California test score data**, estimate the regression below using Nonlinear Least Squares. Report your coefficient estimates and standard errors.

$$TestScore_i = \beta_0(1 - e^{-\beta_1(Income_i - \beta_2)}) + u_i$$

Midterm 2
Due: Wednesday, July 9 by 11:59 pm.
Please submit as a single pdf document on Canvas.
All the data sets are on Canvas.

5. **[Bonus question: 10 points]** Use the **Consumption.xlsx** data on Canvas. We have previously estimated the nonlinear consumption function below using nonlinear least squares in class:

$$C = \alpha + \beta Y^{\gamma} + \varepsilon$$

Where C is the real consumption and Y is the real disposable income. Alternatively, we can assume that the error term has a normal distribution and estimate the nonlinear regression above using the maximum likelihood estimation (MLE) approach. In particular, the MLE approach maximizes the log-likelihood function given by:

$$L(\alpha, \beta, \gamma, \sigma^2) = -\frac{n}{2}\log(\sigma^2) - \log(2\pi) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(C_i - \alpha - \beta Y^{\gamma})^2$$

where σ^2 is the variance of the error term. Using a statistical programming language of your choice, estimate the regression model using the maximum likelihood estimation approach. Your estimates are expected to be similar to those in Table 7.1 of Greene's textbook.
Please submit the following:
a. Your code used to perform the estimation.
b. The output of your estimation, including the estimated parameters and the error variance.