Name:

Student ID#:

Date:

HW4, Chapters 8-10

**Due Friday, April 10, 11:59pm via email (bwjordan2@alaska.edu )**

Instructions:

*This homework is worth 100 points toward your homework grade. These questions draw from chapters 8-10 of the textbook.*

*Part 1 is worth 30 points and should be completed on myeconlab.*

*Part 2 is a set of empirical questions worth 70 points. Do these calculations using R. Write a R script as you work. Write your answers in a Word document. Include in this document any figures, tables, and written answers, print it and attach the pages here. Also paste your R script at the end of this document.*

**Part 1: (30 points)**

In myeconlab, complete the assignment HW 4 Chp 8-10.

**Part 2: (70 Points)**

**Question (1) College distance (35 points)**

The data file CollegeDistance contains data from a random sample of high school seniors interviewed in 1980 and re-interviewed in 1986. In this exercise, you will use these data to investigate the relationship between the number of completed years of education for young adults and the distance from each student’s high school to the nearest four-year college. (Proximity to college lowers the cost of education, so that students who live closer to a four-year college should, on average, complete more years of higher education.) A detailed description is given in College Distance\_Description, also available on the BlackBoard.

Note: *ln(x)* is the mathmatical notation for the natural log of x. The R function however is log(x)

1. Run a regression of ED on Dist, Female, Bytest, Tuition, Black, Hispanic, Incomehi, Ownhome, DadColl, MomColl, Cue80, and Stwmfg80.
   1. If Dist increases from 2 to 3 (that is, from 20 to 30 miles), how are years of education expected to change?
   2. If Dist increases from 6 to 7 (that is, from 60 to 70 miles), how are years of education expected to change?
   3. For both answers, are the results statistically signficiant; based on what evidience? Report 95% confidence intervals.

1. Run a regression of the log of years of education, ln(ED) on Dist, Female, Bytest, Tuition, Black, Hispanic, Incomehi, Ownhome, DadColl, MomColl, Cue80, and Stwmfg80.
   1. If Dist increases from 2 to 3 (from 20 to 30 miles), how are years of education expected to change?
   2. If Dist increases from 6 to 7 (from 60 to 70 miles), how are years of education expected to change?
   3. For both answers, are the results statistically signficiant; based on what evidience? Report 95% confidence intervals.
2. Run a regression of ED on Dist, Dist^2, Female, Bytest, Tuition, Black, Hispanic, Incomehi, Ownhome, DadColl, MomColl, Cue80, and Stwmfg80.
   1. Are dist and dist^2 statistically signficant?
   2. If Dist increases from 2 to 3 (from 20 to 30 miles), how are years of education expected to change?
   3. If Dist increases from 6 to 7 (from 60 to 70 miles), how are years of education expected to change?
3. Do you prefer the regression in (c) to the regression in (a)? Explain.

1. Consider a Hispanic female with Tuition = $950, Bytest = 58, Incomehi = 0, Ownhome = 0, DadColl =1, MomColl =1, Cue80 = 7.1, and Stwmfg = $10.06. Be very careful with these units when you plug them into your new dataset with the predict() function. Double check CollegeDistance\_DataDescription. Remember, you can find the predicited values of a model at particular values doing something like this in R:

YrsPredicted\_linear=data.frame(dist=seq(0,10,1),PredictedYears=predict(m1,newdata=data.frame(dist=seq(0,10,1), female=1, bytest=58, ......)))

This will create a new dataframe with 2 variables. One is ``dist“, that takes values in a sequence: 1,2,3...10. The second variable `` PredictedYears“ are the expected years of education given the coeficient estiamtes from a stored model *m1* and the specific values given by the newdata argument. You can do the same thing for the second order ploynomial model from part (c), I called that one YrsPredicted\_squared.

* 1. Plot the regression relation between Dist and ED from (a) and (c) for Dist in the range of 0 to 10 (from 0 to 100 miles). Describe the similarities and differences between the estimated regression functions.

Hint: You can use the following code to take the predicited data you created (above I called mine YrsPredicted\_linear) to plot the linear model in (a) to the 2nd order polynomial in (c). The argument: type=“l“ makes the plot a line plot, rather than the defualt scatter plot. The function lines adds a line plot on top of whatever plot you have most recently generated.

plot(YrsPredicted\_linear,type="l")

lines(YrsPredicted\_squared)

* 1. Would your answer change if you plotted the regression function for a white non-hispanic male with the same characteristics?
  2. How does the regression function (c) behave for Dist >10? How many observation are there with Dist >10?

1. Add the interaction term DadColl \* MomColl to the regression in (c).
   1. What does the coefficient on the interaction term measure?
   2. Is it significant
   3. How many additional years of education would the following types of person obtain, on average, relative to someone who had neither parent in college:

only dad with college degree:

only mom with college degree:

Both:

1. Does the effect of Dist on ED depends on the family’s income?
   1. Write down a model that could help you answer this question

eg: (Ed = a + b1\*dist....+ b\_x \* Other Controls ). You can group your control variables together in a single variable like this so that the notation is a little cleaner.

* 1. Estimate the model and paste the relevent results here
  2. Using your estimated model, does the the effect of distance on education depend on income?

1. Given the empirical analysis you have done in parts a-g, summarize the effect of distance on years of education.

**Question (2) Traffic crashes (35 points)**

Traffic crashes are the leading cause of death for Americans between the ages of 5 and 32. Through various spending policies, the federal government has encouraged states to institute mandatory seat belt laws to reduce the number of fatalities and serious injuries. In this exercise you will investigate how effective these laws are in increasing seat belt use and reducing fatalities. The data file **Seatbelts** contains a panel of data from 50 U.S. states plus the District of Columbia for the years 1983 through 1997. A detailed description is given in **Seatbelts\_Description**.

“The Effects of Mandatory Seat

Belt Laws on Driving Behavior and Traffic Fatalities,” The Review of Economics and

Statistics, 2003, Vol. 85, No. 4, pp 828-843

1. Plot the relationship between traffic fatalities (fatalityrate) and seat belt usage (sb\_useage), color coded by state.
2. What are some potentially important omitted variables that could bias estimating the effect of seat belt use on traffic fatalities.
3. Estimate the effect of seat belt use on fatalities by regressing *FatalityRate* on

*sb\_useage*, *speed65*, *speed70*, *ba08*, *drinkage21*, ln(*income*), and *age*. Does the estimated regression suggest that increased seat belt use reduces fatalities? Make sure to describe precisely how the fatality rate changes for a change in seat belt use and whether it is statistically significant.

1. Do the results change when you add state fixed effects? Make sure to describe precisely how the fatality rate changes for a change in seat belt use and whether it is statistically significant. Provide an intuitive explanation for why the results changed.

1. Do the results change when you add time fixed effects plus state fixed effects?
2. Which regression specification—(c), (d), or (e)—is most reliable? Explain why.

1. There are two ways that mandatory seat belt laws are enforced: “Primary” enforcement

means that a police officer can stop a car and ticket the driver if the officer observes an occupant not wearing a seat belt; “secondary” enforcement means that a police officer can write a ticket if an occupant is not wearing a seat belt, but must have another reason to stop the car. In the data set, *primary* is a binary variable for primary enforcement and *secondary* is a binary variable for secondary enforcement. Run a regression of *sb\_useage* on *primary*, *secondary*, *speed65*, *speed70*, *ba08*, *drinkage21*, ln(*income*), and *age*, including fixed state and time effects in the regression. Does primary enforcement lead to more seat belt use? What about secondary enforcement?

**Extra credit: (Up to 30 points)**

There are many datasets online related to the recent coronavirus pandemic. For example: <https://covidtracking.com/>. Their data on state level testing, positive and negative cases, deaths are available in csv format: <http://covidtracking.com/api/states/daily.csv>.

The outbreak has many impacts across economic, social, political, and enviornmental demensions.

1. (5 points) create one or two plots of data on coronavirus. Use these plot(s) to tell a story in a paragraph.

(25) points. Find another dataset online that describes some social, policical, or economic demension. Merge these two datasets together. For help merging, see: <https://www.datacamp.com/community/tutorials/merging-datasets-r>

Use your merged dataset and some type of regression model to describe a relationship between the spread of the virus and the social, policical, or economic variable(s). For the full 25 points of extra credit, describe what relationship you chose to study, why the relatition is important, the regression results you obtained, and the results mean. This should be in 1-2 paragraphs.