Homework 4 **Soc 723**
Multiple Regression

These exercises are based on The Effect, chapter 13.

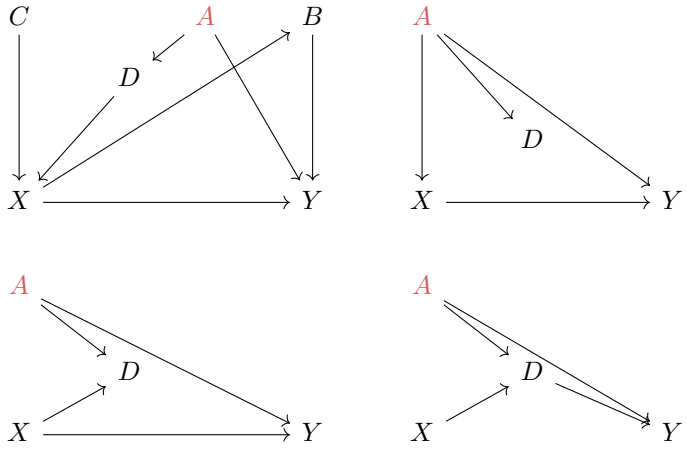
Problem set part 1

1. You've generated some random data X, Y and ε where you randomly generated X and ε as normally distributed data, and then created Y using the formula $Y = 2 + 3X + \varepsilon$. You look at some of the random data you generated, and see an observation with $X = 2$ and $Y = 9$. Let's call that Observation A.

   (a) What is the error for Observation A?

   (b) You estimate the regression $Y = \beta_0 + \beta_1 X + \varepsilon$ using the data you generated and get the estimates $\hat{\beta}_0 = 1.9$ and $\hat{\beta}_1 = 3.1$. What is the residual for Observation A?

2. For each of the DAGs below write the regression equation that you would use to estimate the **total** effect of X on Y, if you think the correct causal diagram is the one below. Perfect identification might not be possible in all cases. **Variables in red are unobserved.** You can name the DAGs by their position: top/bottom left/right.



3. Consider the conventional OLS regression table below, which uses data from 1987 on how many hours women work in paid jobs [1]. In the table, hours worked is predicted using the number of children under the age of 5 in the household and the number of years of education the woman has. Standard errors for each coefficient appear inside parenthesis.

[1] Lee, Myoung–Jae (1995) "Semi–parametric estimation of simultaneous equations with limited dependent variables : a case study of female labour supply", Journal of Applied Econometrics

Annual Hours Worked	**Model 1**	**Model 2**	**Model 3**
(Intercept)	230.018	1256.671	306.553
	(79.671)	(18.046)	(77.975)
Years of Education	72.130		76.185
	(6.232)		(6.09)
Children under 5		-238.853	-251.181
		(19.693)	(19.28)
Num. Obs.	3382	3382	3382
R^2	0.038	0.042	0.084

(a) How many additional hours worked is associated with a one-unit increase of years of education when controlling for number of children?

(b) What is the standard error on the "children under 5" coefficient when not controlling for years of education?

(c) In the third model, what is the predicted number of hours worked for a woman with zero children and zero years of education?

(d) In **the second model**, how many hours worked would the model predict for a woman with 2 children and 10 years of education?

(e) In models 2 and 3, is the coefficient on children under 5 "statistically significantly" different from 0 at the 95% level? How many standard deviations away from 0 are they? And how do we call this statistic?

4. Using the same data as in the question before, we can estimate the model:

$$AnnualHoursWorked = \beta_0 + \beta_1 \text{Edu} + \beta_2 \text{Edu}^2$$

where $\hat{\beta}_0 = 10.145$, $\hat{\beta}_1 = 110.230$ and $\hat{\beta}_2 = -1.581$

(a) What is the relationship between a one-year increase in education and the number of annual hours worked?

(b) What is the relationship between a one-year increase in education and annual hours worked if the current level of education is 16?

(c) Is the relationship between education and hours worked getting more or less positive for higher values of education?

(d) What would be one reason not to include a whole bunch of additional powers of education in this model (Edu^3, Edu^4, and so on)?

5. The following table uses the same data from the previous questions, but this time all of the predictors are binary. The first model predicts working hours using whether the family owns their home, and the second uses the number of children under 5 again, but this time treating it as a categorical variable.

Annual Hours Worked	**Model 1**	**Model 2**
(Intercept)	1101.313	1242.904
	(27.168)	(18.839)
Homeowner	50.174	
	(32.923)	
1 Child under 5		-158.164
		(35.800)
2 Children under 5		-526.006
		(50.779)
3 Children under 5		-773.412
		(113.394)
4 Children under 5		-923.904
		(357.031)
Num. Obs.	3382	3382
R^2	0.001	0.044

(a) Interpret the coefficient on "Homeowner". Is this coefficient "statistically significantly" different from 0 at the 95% level?

(b) According to the model, how many fewer hours do people with 4 children under the age of 5 work than people with 3 children under the age of 5?

(c) From this table alone can we tell whether there's a "statistically significant" difference in hours worked between having 2 children and having 3? How could we test that?

6. Consider the below regression table, still using the same data on annual number of hours worked.

	Hours Worked Model 1	**log(Hours Worked)** Model 2	**Hours Worked** Model 3
(Intercept)	-244.147	6.243	-954.379
	(143.761)	(0.164)	(180.681)
Homeowner	682.992	0.897	
	(172.921)	(0.196)	
Education	110.073	0.067	
	(11.558)	(0.013)	
Homeowner x Education	-53.994	-0.063	
	(13.738)	(0.015)	
log(Education)			832.347
			(71.684)
Num. Obs.	3382	2487	3376
R^2	0.043	0.015	0.038

(a) In Model 1, what is the relationship between a one-unit increase in Education and annual hours worked?

(b) Interpret the coefficient on Homeowner × Education in Model 1.

(c) Interpret the coefficient on Education in Model 2. Note that the dependent variable is log(Annual Hours Worked).

(d) Interpret the coefficient on log(Education) in Model 3 using percentage change in education.

(e) Why do you think the sample size is different in model 2?

7. Which of the following is the most accurate definition of autocorrelation in an error term?

(a) When error terms are correlated within the same (auto-) group, for example when test scores being more similar within classrooms than between them.

(b) When error terms are correlated across time, such that knowing the error term in one period gives us some information about the error term in the next period.

(c) When a variable that's measured across time has a trend in it, for example trending upwards or trending downwards.

(d) When a sandwich estimator is used to allow for correlation across a time series.

8. You have run an OLS regression of Y on X, and you would like to figure out whether it would be a good idea to use heteroskedasticity-robust standard errors. Which of the following would help you figure this out? Select all that apply.

(a) Creating a plot with Y on the y-axis and X on the x-axis, and a line reflecting the predicted values of the regression, and seeing if the predicted values change over the range of X.

(b) Creating a plot with Y on the y-axis and X on the x-axis, and a line reflecting the predicted values of the regression, and seeing if the spread of the values around the predicted values change over the range of X.

(c) Creating a plot with Y on the y-axis and Z on the x-axis (where Z is not included in your model), and a line reflecting the predicted values of the regression, and seeing if the spread of the values around the predicted values change over the range of Z.

(d) Checking if the R^2 value of the regression is particularly low.

(e) Asking whether Y is continuous or binary.

9. Political pollsters gather data by contacting people (by phone, knocking on their door, internet ads, etc.) and asking them questions. A common problem in political polling is that different kinds of people are more or less likely to respond to a poll. People in certain demographics that have historically been mistreated by pollsters, for example, might be especially unlikely to respond, and so the resulting data will not represent those groups well. If a pollster has information on the proportion of each demographic in a population, and also the proportion of each demographic in their data, what tool from Chapter 13 can they use to help address this problem, and how would they apply it?

10. Which of the following is an example of measurement error where we can tell that the measurement error is non-classical?

    (a) You're doing research on unusual sexual practices. You ask people whether they've ever engaged in these practices, which many people might prefer to keep secret, even if they've actually done them.

    (b) You're measuring temperature, but because the thermometer is imprecise, it only measures the actual temperature within a few degrees.

    (c) You're looking at the relationship between athleticism and how long you live. As your measure of how athletic someone is, you use their time from running a kilometer when they were age 18, since you happen to be studying a country where nearly everyone had to do that before leaving school.

Problem Set Part 2

Load the "dengue.csv" file. Documentation on the variables is available here.

1. Run an OLS regression using average humidity and temperature to predict whether dengue was observed in the area, and plot a summary table of the results.

2. Our dependent variable is binary, and we're getting predictions below zero, which we might not want. Rerun the regression from the previous question but as a logit model, report and interpret the marginal effects of both slope coefficients.

3. Now let's say we're directly interested in the relationship between temperature and humidity. Run an OLS regression of humidity on temperature. Calculate the residuals of that regression, and then make a plot that will let you evaluate whether there is likely heteroskedasticity in the model. Rerun the model with heteroskedasticity-robust standard errors. Show both models, and say whether you think there is heteroskedasticity. Note: remove NAs for the variable `humid` on the data before running the model so the residuals line up properly.

4. In the graph in the last problem you may have noticed that for certain ranges of temperate, the errors were clearly nonzero on average. This can indicate a functional form problem. Run the model from the previous question again (with heteroskedasticity-robust standard errors), but this time use the logarithm of humidity in place of humidity. Add a sentence interpreting the coefficient on temperature.

5. Bonus question: figure out how I decided on a form where you log humidity and keep temperature linear.