**Samuel Snelson**

**SOC 723 Spring 2023**

**Homework for Chapter 13: Regression**

1. You’ve generated some random data , , and where you randomly generated and as normally distributed data, and then created using the formula . You look at some of the random data you generated, and see an observation with and . Let’s call that Observation A.
   1. What is the *error* for Observation A?
      1. NHK differentiates between error and residual. The residual is simply the difference between observed and predicted data. This is an estimate of the error, or the difference between the observed data and the predictions of the *true* model. In real situations, we can’t actually say anything about the error directly but use residuals as an estimate with parity depending on factors like sampling variation. In this scenario, the outcome variable is simulated so model residuals are equivalent to errors.
      2. The model Y = 2 + 3X predicts (2,8). With an observation at (2, 9), there is an estimated error on this observation of 1 unit of Y.
   2. You estimate the regression using the data you generated and get the estimates . What is the *residual* for Observation A?
      1. With an estimated model of the outcome, we now can compute residuals as an estimate of errors.
2. Write the regression equation that you would use to estimate the effect of on , if you think the correct causal diagram is the one below. Assume you can measure all the variables in the diagram.

Diagram

Description automatically generated

* 1. Assuming this is the correct model, there are two ways to identify the effect of X on Y. We could isolate the purely exogenous variation in X due to C. If we think of X as an experimental treatment and C as random selection into treatment, backdoors are closed off because comparisons are directly between levels of C (although it may be more complicated than the simple case of random assignment). As the estimation of this goes, I am not quite sure yet how to do this. Though, we would be interested in regressing Y on X conditioning for the levels of C. Another way to estimate the effect of *X* and *Y* would be to close open backdoors from *X* to *Y*. There are three of these “bad” paths that all pass through *A* and *B*. To estimate the effect of *X* on *X* on *Y*, we can condition on *A* and *B* (noting that here C is antecedent and irrelevant for the estimation if we believe backdoors have been closed)
  2. I presume NHK is asking about the latter (regular regression), but in either case it would be efficient to adjust for A and B, so we end up with the following equation:

1. You use regression to estimate the equation and get an estimate of and the standard error .
   1. Interpret, in a sentence, the coefficient .
      1. A one unit change in X is (linearly) associated with a 1.3 unit change in Y.
   2. Calculate whether is statistically significantly different from 0 at the 95% level. (more technical detail you may not need: do a two-sided test, and assume the sample size is effectively infinite)
      1. With infinite sample size, we can just compute the z-statistic from the ratio of the coefficient and its standard error.
      3. What does this Z statistic tell us? This statistic is located on the normal Z distribution with N(0,1). If we are testing whether a parameter is significantly different from 0 at the 95% confident level, then the parameter would have to exceed the Z score of 1.96 (or -1.96). 1.96 (and -1.96) refer to the portion of the Z distribution with a probability density of 2.5%. Using a two-tailed test, we consider that a statistic may be positive or negative and take the 2.5% corresponding to 1.96 and -1.96 together. 2.3 is greater than 1.96. This tells us that the probability of observing a test statistic as or more extreme than 2.3 (>= 2.3 and <= -2.3), assuming that the *true* test statistic was 0, is less than 0.05.
   3. Whatever your answer to part b, what does it mean to say that this coefficient is statistically significantly different from 0?
      1. See above answer. The logic of this type of null hypothesis significant testing is that if we assume a world is which the null parameter is true (coefficient is 0), then it would be highly unlikely that we would observe the parameter that we have (2.3). In such situations, this serves as evidence that the imagined world of the null parameter is probably not true.
2. Consider the below conventional OLS regression table, which uses data from 1987 on how many hours women work in paid jobs.[[1]](#footnote-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.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Annual Hours Worked  (1)** | **Annual Hours Worked  (2)** | **Annual Hours Worked  (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 |
| R2 | 0.038 | 0.042 | 0.084 |
| \* p < 0.1, \*\* p < 0.05, \*\*\* p < 0.01 | | | |

* 1. How many additional hours worked is associated with a one-unit increase of years of education when controlling for number of children?
     1. Conditioning on the number of children under 5 in a household, a one year increase in a women’s education is associated with about a 76 hour increase in the number of hours a women works annually.
  2. What is the standard error on the “children under 5” coefficient when not controlling for years of education?
     1. 19.693
  3. In the third model, what is the predicted number of hours worked for a woman with zero children and zero years of education?
  4. How many observations are used in each of the three regressions?
     1. 3382
  5. Is the coefficient on “children under 5” statistically significantly different from 0 at the 95% level?
     1. Yes

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

We can note immediately that this model is nonlinear with a parabolic form opening down. This means that the relationship between education and hours worked starts out with more education increasing annual hours worked. However, the relationship reaches an apex meaning that as a women becomes more and more educated, their annual hours ceases to increase and starts to decrease.

* 1. What is the relationship between a one-year increase in and ? (hint: your answer will not just be a single number, it will still include a term)
     1. To identify the effect of the number of years of education on annual hours worked, we can take the derivative of the second and third parts of the equation because it is nonlinear. We can simply apply the power rule. Thus, a one-year increase in the number of years of education a woman has is associated with an increase in the annual working hours of:
  2. What is the relationship between a one-year increase in and if the current level of is 16?
     1. The effect of an increase in the number of years of education of 16 to 17 on annual working hours is given by the difference in the effects of education on hours worked for a woman with 17 relative to 16 years of education.
        1. Thus, the effect of a women going from 16 to 17 years is about 58 additional hours worked annually. Given education is still positively related to hours worked this far along the education distribution, that tells us that the purpose of the quadratic form may have been primarily to allow for diminishing returns of education on hours worked (rather than a linear increase).
  3. Is the relationship between and getting more or less positive for higher values of ?
     1. As stated previously, the effect of education on hours worked decreases the higher a women’s education is.
  4. What would be one reason *not* to include a whole bunch of additional powers of in this model ( and so on)
     1. One clear reason not to include a bunch of higher order terms in the model is that it would likely be overfit. As scientists (so we say), we want to capture the actual real world phenomena when we are modeling data on some social process. When we have higher order terms, we are allowing the model to be more sensitive to the sampled data. But there is a cost for allowing too much sensitivity. There may be irregular features in our data that do not actually reflect the phenomena at the population level. When we are concerned with inference (rather than prediction), we want to balance the model fit to the data with consideration of out-of-sample predictive accuracy.

1. The following table uses the same data from question 4, 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 (1)** | **Annual Hours Worked (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 |
| R2 | 0.001 | 0.044 |
| \* p < 0.1, \*\* p < 0.05, \*\*\* p < 0.01.  In model (2), “zero children under 5” is the reference category. | | |

* 1. Interpret the coefficient on “Homeowner”
     1. This coefficient refers to the average difference in the number of hours a woman works annually between women who own their home and who do not. More intuitively, we say that women owning their home is associated with working about 50 more hours annually than women who don’t own their home.
  2. On average, 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?
     1. This difference in the predicted number of hours worked is given by the difference of the coefficients for 4 children relative to 3 children.
        1. Conditioning on homeownership, women who have 4 children under 5 years old are predicted to work about 150 hours less annually than women who have 3 children under 5 years old.
  3. From this table alone can we tell whether there’s a statistically significant difference in hours worked between having 2 children and having 3? What additional test would we need to perform?
     1. We cannot tell just from this information. These significance tests refer to the difference in the annual hours worked for women with 1-4 children under 5 years old relative to women with zero children under 5 years old. We would have to perform the same tests for each interval (0 to 1, 1 to 2, 2 to 3…) to see these differences.

1. Consider the below regression table, still using the same data as in 4-6.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Annual Hours Worked  (1)** | **log(Annual Hours Worked)  (2)** | **Annual Hours Worked  (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 |
| R2 | 0.043 | 0.015 | 0.038 |
| \* p < 0.1, \*\* p < 0.05, \*\*\* p < 0.01 | | | |

* 1. In Model 1, what is the relationship between a one-unit increase in Education and annual hours worked?
     1. Because we have an interaction term between homeownership and education in the first model, the effect of education involved both the term just for education and the interaction term.
     2. A one-year increase in the education a woman has is associated with the following change in the annual hours worked:
        1. What does this mean? Because we are modeling interaction, the effect of education depends on homeownership (and conversely the effect of homeownership depends on education). To put this another way, the effects of education may be different for homeowners relative to women who don’t own.
  2. Do annual hours worked rise more quickly for homeowning families or non-homeowning families? Is the difference between the two statistically significant at the 95% level?
     1. Based on the homeownership coefficient in the first model, there is a considerable increase in the estimated number of hours worked annually for women who own their own (conditioning on education, the effect of homeownership in the first model is an additional hours worked annually. The coefficient for homeownership is statistically significant at the 95% alpha level (but to understand the relationship between homeownership and hours worked, we have to look at both the term alone and the interaction term).
  3. Interpret the coefficient on Homeowner x Education in Model 1.
     1. I find a helpful way to think about interaction coefficients as being a contribution to their respective main variables. Looking at the homeownership, education, and interaction terms together, they tell us that for women with few years of education, homeownership is associated with more annual working hours. As a women’s education increases, those who own homes work fewer annual hours than women who don’t own a home.
  4. Interpret the coefficient on Education in Model 2. Note that the dependent variable is *log* annual hours worked.
     1. There are pretty much two ways to interpret coefficients with log models. You can try to think about change on the log scale or convert back to the raw scale by exponentiating (like with logit models).
     2. A one year increase in education is associated with a change in the log of annual hours worked. If we want to interpret in the raw scale (more sensibly) in terms of hours, we can exponentiate this expression which gives us . Among women who own their home, a one year increase in education is associated with a .04% increase in annual hours worked (quite small as we mentioned earlier – the diminishing effect of education is apparent). Among women who don’t own a home, a one year increase in education is associated with about a 7% increase in annual hours worked.
  5. Interpret the coefficient on log(Education) in Model 3, beginning with “a 10% increase in Education…”
     1. A 10% increase in education is associated with about an 83 hour increase in annual hours worked.
  6. Why do you think the sample sizes are different in each of the three models? The only thing that really changed was the addition of the logarithms…
     1. Because the natural log of zero is undefined, there may be women who do not work outside the home which have been dropped from the analytical sample. I am not so sure about this reasoning for the decreased analytical sample where education was logged because I’m not so sure there would be any people with zero years of education.

1. Which of the following is the most accurate definition of *autocorrelation* in an error term?
   1. When error terms are correlated within the same (auto-) group, for example when test scores being more similar within classrooms than between them
   2. 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
   3. When a variable that’s measured across time has a trend in it, for example trending upwards or trending downwards
   4. When a sandwich estimator is used to allow for correlation across a time series

Option A refers to spatial autocorrelation, where residuals would be clustered by classrooms. This may be because of differences in classroom environments or teacher effectiveness. Students may be systematically worse or better in certain classrooms. Option B refers to temporal autocorrelation, where residuals at one time are correlated with residuals at a later time (I suppose without multilevel modeling of within person differences, this would be problematic if repeated measurements were assumed to be independent and identity drawn).

1. You have run an OLS regression of on , 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**.
   1. Creating a plot with on the y-axis and 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 .
   2. Creating a plot with on the y-axis and 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 (if the heteroscedasticity is evident, then this will indicate that because the spread of the outcome across the x distribution is where heteroscedasticity comes from. The common residual plot is just this plot reoriented where the y-axis is the residuals).
   3. Creating a plot with on the y-axis and on the x-axis (where 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
   4. Checking if the value of the regression is particularly low
   5. Asking whether is continuous or binary (if you are doing an OLS regression on a binary outcome – an LPM – then heteroscedasticity is guaranteed to be maximum at 0.5. Using robust errors addresses this issue).
2. 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?
   1. Weighting would be a useful approach to resolve this issue of disproportionate representation. Commonly weights are applied to increase the weight of certain observations in the estimation of effects. If we see a massive difference in a small subset of a sample, then this may be weighted less than a more modest difference in a majority subset of a sample. However, weights can also be applied in an inverse fashion to invert under and over-representation – this is called inverse sample probability weights.
   2. If we wanted to apply an inverse sample probability weight in a situation of disproportionate distribution of sample categories, we could simply multiply desired sample statistics by the inverse of the sample proportion of each group (e.g., if a group made up 15% of the total sample, their inverse sample probability weight would just be 1/0.15 = 6.67.
3. Which of the following is an example of measurement error where we can tell that the measurement error is *non-classical*?

Based on NHK’s description of non-classical measurement error, I think about it in terms of there being an open backdoor from the predictor being used and unobserved variables in the error term. If the error term influenced the predictor, this is non-classical and presents some challenges. If it is classical measurement error, there is no backdoor from the predictor to the error – just random noise.

* 1. You’re doing research on unusual sexual practices. You ask people whether they’ve ever engaged in these weird practices, which many people might prefer to keep secret, even if they’ve actually done them. (This is a good example of likely non-classical measurement error because a person very well might give an inaccurate answer not randomly but because of the measurement itself).
  2. You’re measuring temperature, but because the thermometer is imprecise, it only measures the actual temperature within a few degrees (if the imprecision is random with respect to the measurement, then this is classical measurement error – error is independent of the measurement).
  3. 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. (For this to be classical measurement error, the errors introduced by the difference in athleticism at the time the person ran and their general athleticism throughout their life that would influence life expectancy would have to be random. Some people ran well at 18 and kept athletic habits and some didn’t – randomly. Some people well poorly at 18 and improved their habits or kepts poor habits – randomly. I’m not sure that there would be systematic differences, so we’ll go with classical).

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, 10(2), April–June, 187–200. [↑](#footnote-ref-1)