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?

=9-2-3\*2=1

* 1. You estimate the regression using the data you generated and get the estimates . What is the *residual* for Observation A?

Given our model is Y=1.9+3.1X, Y turns to be 8.1 when X=2. The residual is thus (9-8.1)=0.9.

1. 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

Y=β0+β1X+β2A+β3B+ℇ1

1. You use regression to estimate the equation and get an estimate of and the standard error .
   1. Interpret, in a sentence, the coefficient .

One-unit increase in X is linearly associated with a 3-unit increase in Y.

* 1. 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)

/=3/1.3≈2.3

The t-statistic above 1.96 (the 97.5th percentile) which means the is statistically significantly different from 0 at the 95% level.

* 1. Whatever your answer to part b, what does it mean to say that this coefficient is statistically significantly different from 0?

If the coefficient is statistically significant different from 0 with α = .05, our estimate is below the 2.5th percentile or above the 97.5th, and has a 5% chance or less of occurring if the null value 0 is true, which makes us reject the null and think it’s unlikely to be 0, i.e., there’s a non-zero relationship there.

1. 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?

76.185.

* 1. What is the standard error on the “children under 5” coefficient when not controlling for years of education?

19.693.

* 1. In the third model, what is the predicted number of hours worked for a woman with zero children and zero years of education?

306.553.

* 1. How many observations are used in each of the three regressions?

3382.

* 1. Is the coefficient on “children under 5” statistically significantly different from 0 at the 95% level?

Yes, since the coefficient on “children under 5” has three stars indicating its p-value is under 0.01, i.e., its probability of being as far away from the null hypothesis value (or farther) is below 0.01 and thus below 0.05.

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

A one-year increase in *YearsEducation* is associated with a (110.230-3.162 *YearsEducation*)-hour change in .

* 1. What is the relationship between a one-year increase in and if the current level of is 16?

A one-year increase in *YearsEducation* is associated with a 59.638-hour increase in .

* 1. Is the relationship between and getting more or less positive for higher values of ?

The relationship between and gets less positive for higher values of .

* 1. What would be one reason *not* to include a whole bunch of additional powers of in this model ( and so on)

The model becomes much harder to interpret. In many cases, by adding more and more polynomial terms can make the model more complex without actually improving its fit, instead lead to “overfitting,” where a too-flexible model winds up bending in strange shapes to chase individual observations and noise rather than the overall relationship in the data, producing a worse model.

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”

On average, the annual working hours for women in paid jobs are 50.174 higher when the family owns their home than when the family does not own their home.

* 1. 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?

Y4=-158.164x1-526.006x2-773.412x3-923.904\*1+1242.904=-923.904+1242.904=319,

Y3=-158.164x1-526.006x2-773.412\*1-923.904x4+1242.904= -773.412+1242.904= 469.492, ∆D= 469.492-319=150.492

So on average, people with 4 children under the age of 5 work 150.492 fewer hours than people with 3 children under the age of 5.

* 1. 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?

We cannot tell whether there’s a statistically significant difference in hours worked between having 2 children and having 3 from this table alone. We need to do a hypothesis test on a new model Annual Hours Worked = β0+ β1\*0 Children under 5+ β2\*1 Children under 5 + β3\*3 Children under 5 + β4\*4 Children under 5 with which we set the “2 Children under 5” as a reference category and see if the coefficient on “3 Children under 5” is statistically significantly different from 0 at the 95% level via the t-statistic.

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?

Controlling for other variables, a one-unit increase in Education is associated with a 110.073-hour increase in women’s Annual Working Hours for paid jobs.

* 1. Do annual earnings (or hours worked?) rise more quickly for homeowning families or non-homeowning families? Is the difference between the two statistically significant at the 95% level?

Annual hours worked or annual earnings (assuming annual earnings is positively associated with the annual hours worked) rises more quickly for homeowning families compared with non-homeowning families. The difference between the two is statistically significant at the 95% level since the p-value of the coefficient on Homeowner in model 1 is under 0.01, i.e., its probability of being as far away from the null hypothesis value (or farther) is below 0.01 and thus below 0.05.

* 1. Interpret the coefficient on Homeowner x Education in Model 1.

The coefficient on Homeowner x Education in Model 1 is how much weaker (-53.994) the effect of Education on Annual Working Hours for paid jobs gets when Homeowner increases by one unit, going from Homeowner = 0 to Homeowner = 1, i.e., difference in the effect of Education on Annual Working Hours for paid jobs between homeowning families and non-homeowning families.

* 1. Interpret the coefficient on Education in Model 2. Note that the dependent variable is *log* annual hours worked.

The coefficient on Education in Model 2 means a 1-unit increase in Education is associated with a 0.067-hour change in log(annual hours worked), which then means a 6.7% increase in annual hours worked.

* 1. Interpret the coefficient on log(Education) in Model 3, beginning with “a 10% increase in Education…”

A 10% increase in Education is associated with an 83.2347-unit increase in annual hours worked.

* 1. 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…

Because logarithm can’t handle values of zero, log(0) is undefined, the logarithm transformation we made left out observations with zeroes in them, for example, when the education is 0.

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
2. ~~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~~**~~. (2)~~
   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
   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
3. 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?

We can use inverse-sampling-probability weights to help address this problem by weighting each individual by the inverse of the probabilities of them being included in the sample data, i.e., 1/[(the proportion of the demographic group in the data\*sample size)/(the proportion of the demographic group in a population\*population size)]. By comparing the resulting proportions of demographic groups in the population against their data values, we can get the ratio of weights we will give to different demographic groups.

1. Which of the following is an example of measurement error where we can tell that the measurement error is *non-classical*? a
   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.
   2. You’re measuring temperature, but because the thermometer is imprecise, it only measures the actual temperature within a few degrees
   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.
2. See the Chapter 13 Coding Homework

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)