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?

Answer: Y’= 2+3\*2=8; Error=Y-Y’=9-8=1.

The error for observation A is 1.

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

Answer: Y’=1.9+3.1\*2=8.1; Residual=Y-Y’=9-8.1=0.9

The residual is 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.

Answer: Y=  
Diagram

Description automatically generated

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

Answer: Controlling for the other variables, a one-unit change in X is linearly associated with a 3-unit change 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)

Answer: Based the pictures showing the normal distribution with mean of 0 and sd of 1.3 and the calculation of the 2.5% and 97.5%, the 2.5th percentile of the normal distribution is about -2.55 and the 97.5th percentile of the normal distribution is about 2.55. Obviously, our estimate, 3, is the right side of the 97.5 percentile, 2.55. Besides, the estimate, 3, is around at the 98.95th percentile of the normal distribution, which means that something as far from 0 as 3 happens around 2.1% ((100-98.95)\*2)of the time, lower than 5%. So we could say that is statistically significantly different from 0 at the 95% level.

图形用户界面, 应用程序

描述已自动生成

图表, 折线图

描述已自动生成

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

Answer: The coefficient is statistically significantly different from 0 means that with the rejection value of α, 0.05, the estimated coefficient is occurring less than 5% with the theoretical normal distribution (mean=0, sd=1.3), against a null that the coefficient is 0.

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?

Answer: About 76.185 additional hours worked is associated with a one-unit increase of years of education when controlling for number of children.

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

Answer: The standard error is about 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?

Answer: The predicted number of hours worked for a woman with zero children and zero years of education is about 306.553.

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

Answer: 3382 observations.

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

Answer: Yes, the coefficient on “children under 5” is statistically significantly different from 0 at 95% level in model 2 and model 3. (Three stars)

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)

Answer:

A one-year increase in is associated with a ()-unit change in

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

Answer: When YearsEducation is 16, is equal to 59.638. So, a one-year increase in *YearsEducation* is associated with a 59.638 unit change in *AnnualHoursWorked.*

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

Answer: Based on the derivative of the model and the graph below, we could know that, as the value of YearsEducation increases, the relationship first gets less positive (still positive but close to 0), and then when the value of Years Education is equal or higher than 35 (110.230/3.162=34.86), the relationship gets more negative.

图表, 散点图

描述已自动生成

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

Answer: Adding a whole bunch of additional powers of YearsEducation in this model might lead to “overfitting” problem where a too-flexible model will try to fit noise in the data and produce a worse model. And the “overfitting” model might not be useful to be applied to the populations since it is “overfitting” with the sample.

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”

Answer: The coefficient on “Homeowner” means that on average, the number of hours women work in a paid job when the family owns their home is 50.174 higher 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?

Answer: On average, people with 4 children under the age of 5 work for 150.480 (923.904-773.412=150.492) hours fewer 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?

Answer: This table alone cannot tell us whether there’s statistically significant difference in hours worked between having 2 children and having 3. We could take the women with 2 children as the reference category, and take the hypothesis testing again. Then the stars of the coefficient of 3 children could tell us whether there’s a statistically significant difference in hours worked between having 2 children and having 3 children. If we want to know about the overall categorical variable has statistically significant effect, we could also take the joint F test.

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?

Answer: A one-unit increase in Education is associated with a (110.073-53.994\*Homeowner)-unit change in annual hours worked. In other words, a one-unit increase in Education is associated with a 56.079-unit (110.073-53.994\*1) change in annual hours worked for women whose family owns their home; while for women whose family does not own their home, a one-unit increase in Education is associated with a 110.073-unit(110.073-53.994\*0) change in annual hours worked.

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

Answer: If annual earnings is associated with annual hours worked and education years, then annual earnings rise more quickly for non-homeowning families than for homeowning families. Based on the number of stars of the coefficient of Homeowner x Education which is 3 stars, we could say that the difference between the two is statistically significant at the 95% level.

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

Answer: The coefficient on Homeowner x Education in Model 1 means that the effect of Education on AnnualHoursWorked is 53.994 more negative for people whose family owns their home than for people whose family does not own their home.

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

Answer: In model 2, the coefficient on Education means that a one-unit increase in Education is associated with a (0.067-0.063\*Homeowner)-unit change in log annual hours worked. In other words, a one-unit increase in Education is associated with a (6.7-6.3\*Homeowner)-unit change in annual hoursworked.

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

Answer: A 10% increase in Education is associated with a 83.2347-unit change 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…

Answer: Because logarithms cannot handle values of zero. So when the logged variables have zero values, when have to drop out the observations with zero values of the logged variables.

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**.
   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?

Answer: We could use sample weights to address this problem. For example, we can first calculate the inverse probability of the proportion of each group in data over the proportion of each group in population. Then we could weight each individual by the inverse of those probabilities.

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

Answer: A. (For C, it could also be the non-classical error when the time for running is reported by themselves. The person who is less athletic might be possible to report the less time they actually use for running a kilometer. Then the error is related to the latent variable.)

* 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.

1. 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)