

Examples of Questions on Regression Analysis:

1. Suppose that a score on a final exam depends upon attendance and unobserved factors that affect exam performance (such as student ability). Then,

$$score = \beta_0 + \beta_1 attendance + \epsilon.$$

When would you expect this model to satisfy the assumption $E(\epsilon | attendance) = 0$? (That is, when would you expect the conditional expectation of the unobserved error term to be zero?)

Answer: you would expect this to be the case if the unobserved variables that can affect exam performance (such as student ability) are NOT correlated with attendance. This seems pretty unlikely, though.

2. Let kids denote the number of children ever born to a woman, and let educ denote years of education for this woman. A simple model relating fertility to years of education is:

$$kids = \beta_0 + \beta_1 educ + \epsilon.$$

- a. What kinds of factors are contained in ϵ ? Are these likely to be correlated with the level of education?
- b. Will a simple regression analysis uncover the ceteris paribus effect of education on fertility? Explain.

Answer:

a. Other factors that might be contained in the error term would be age of the mother, income level, religion. Surely age and income would be highly correlated with education.

b. If you ran the above regression, you wouldn't be controlling for any other effects (there are no other effects in the model: no other RHS variables) so you would NOT be looking at the ceteris paribus effect of education on fertility. A simple regression would tell you the OVER-ALL effect of education on kids (controlling for nothing else at all).

3. In the estimated linear consumption function: $\hat{c}ons = \hat{\beta}_0 + \hat{\beta}_1 inc$ the (estimated) marginal propensity to consume (MPC) out of income is simply the slope, $\hat{\beta}_1$ and the average propensity to consume out of income (APC) is given by $\hat{c}ons/inc = \hat{\beta}_0/inc + \hat{\beta}_1$. Using 100 families randomly chosen and taking their annual income and consumption data (both measured in dollars) the following equation is obtained:

$$\hat{c}ons = -124.84 + 0.853 inc \quad R^2 = 0.692, \quad N = 100$$

- a. Interpret the intercept in this equation, and comment on its sign and magnitude.
- b. What is the prediction consumption when family income is \$30,000?

Answer:

- a. xxx
- b. predicted consumption = $-124.84 + 0.853(30,000)$

4. The OLS fitted line explaining college GPA in terms of high school GPA and ACT score is estimated as:

$$colGPA = 1.29 + 0.453hsGPA + 0.0094ACT$$

If the average high school GPA is about 3.4 and the average ACT score is about 24.2, what is the average college GPA in the sample?

Answer: $1.29 + 0.453(3.4) + 0.0094(24.2) = 3.06$

5. Suppose there are two candidates for office and we wanted to explain the share of votes that candidate A gets and we model it as:

$$voteshareA = \beta_0 + \beta_1 expendA + \beta_2 expendB + shareA + \epsilon$$

Where $expendA$ is the campaign expenditures spent by candidate A (and $expendB$ is the campaign expenditures of candidate B), and $shareA = expendA / totalexpend$ where $totalexpend$ is the TOTAL amount of campaign expenditures put out by both candidates. Does this model violate the perfect multicollinearity assumption?

Answer: No. There is no perfect linear relationship between the explanatory variables. (There is, however, a perfect NON-LINEAR relationship between the explanatory variables ... but that's okay as far as OLS is concerned.)

6. Suppose that you postulate a model explaining final exam score in terms of class attendance. Thus, the dependent variable is final exam score, and the key explanatory variable is number of classes attended. To control for student abilities and efforts outside the classroom, you include among the explanatory variables cumulative GPA, SAT score, and measures of highschool performance. Someone says, "You cannot hope to learn anything from this exercise because cumulative GPA, SAT score and highschool performance are likely to be highly collinear." What should be your response?

Answer: You're interested in how class attendance affects final exam scores. Even if GPA, SAT score and highschool performance are highly collinear (and therefore might have low t-statistics because OLS can't sort out their relative contribution to

explaining the variation in final exam score), that will NOT affect the OLS estimator on class attendance. The OLS estimator will still be unbiased, as will the variance on the coefficient estimator. If we were to DROP those variables and they were correlated with attendance, then we'd really be in trouble: we would have omitted variable bias.

7. Data on working men was used to estimate the following equation:

$$educ = 10.36 - 0.094sibs + 0.131meduc + 0.210feduc \quad N = 722, R^2 = 0.214$$

were educ = years of schooling, sibs = number of siblings, meduc = mother's years of schooling, feduc = father's years of schooling.

- a. Does sibs have the expected effect? Explain. Holding meduc and feduc fixed, by how much does sibs have to increase to reduce predicted years of education by one year? (A non integer answer is acceptable here.)

The estimated model shows that as the number of siblings is negatively related to the education level of a working man. Holding the mother and father's levels of education fixed, a 10.64 increasing in siblings will reduce the education level by 1 year.

- b. Discuss the interpretation of the coefficient on medec.

There is a positive relationship between a mother's level of education and her son's level of education. For every additional year of mother's education, holding the father's level of education fixed and the number of siblings, a son's education will go up by 0.13 years.

- c. Suppose that Man A has no siblings, and his mother and father each have 12 years of schooling. Man B has no siblings, and his mother and father each have 16 years of schooling. What is the predicted difference in years of education between A and B?

Predicted difference between Man B and Man A's education level:

$$0.131(16) + 0.21(16) - 0.131(12) - 0.21(12) = 4(0.131) + 4(0.21) = \text{some number.}$$

8. The following model is a simplified version of the multiple regression model used by Biddle and Hamermesh (1990) to study the tradeoff between time spent sleeping and working and to look at other factors affecting sleep:

$$sleep = \beta_0 + \beta_1 totwrk + \beta_2 educ + \beta_3 age + \epsilon$$

where sleep and totwrk (total work) are measured in minutes per week and educ and age are measured in years.

- a. If adults trade off sleep for work, what is the sign of β_1 ?

If there is a sleep-work trade off, then the sign on β_1 should be negative.

- b. What signs do you think β_2 and β_3 will have?

β_2 might be negative: the more education you have, the more demanding your job, the less sleep you get...although this could go the other way as well.

β_3 is probably negative: the older you get, the less sleep you need (and get!).

- c. Suppose the following equation is estimated:

$$sleep = 3638.25 - 0.148totwrk - 11.13educ + 2.20age \quad N = 706, R^2 = 0.113$$

If someone works for five more hours per week, by how many minutes is sleep predicted to fall? Is this a large trade-off?

Holding all other things constant, sleep will fall by $5(0.148)(60)$ minutes.

- d. Discuss the sign and magnitude of the estimated coefficient on educ.

More education, less sleep. Very depressing. 4 years of college means 44 minutes less sleep (approximately) compared to a high school graduate.

- e. Would you say totwrk, educ, and age explain much of the variation in sleep? What other factors might affect the time spent sleeping? Are these likely to be correlated with totwrk?

Only 11.3% of the variation in minutes of sleep is explained by the model. Other factors: NUMBER OF CHILDREN. Almost surely this would be correlated with totwrk.

9. The median starting salary for new law school graduates is determined by:

$$\log(salary) = \beta_0 + \beta_1 LSAT + \beta_2 GPA + \beta_3 \log(libvol) + \beta_4 \log(cost) + \beta_5 rank + \epsilon$$

where LSAT is the median LSAT score for the graduating class, GPA is the median college GPA for the class, libvol is the number of volumes in the law school library, cost is the annual cost of attending law school, and rank is a law school ranking (with rank = 1 being the best).

- a. Explain why we expect $\beta_5 \leq 0$.

Better ranked school (better school), higher wage. Rank is measured “backwards” with 1 being the best so we’d expect a negative relationship.

- b. What signs do you expect for the other slope coefficients? Justify your answers.
- c. Suppose the following equation is estimated:

$$\log(\text{salary}) = 8.34 + 0.0047\text{LSAT} + 0.248\text{GPA} + 0.095\log(\text{libvol}) + 0.038\log(\text{cost}) - 0.0033\text{rank}$$

$$N = 136, R^2 = 0.842$$

What is the predicted ceteris paribus difference in salary for schools with a median GPA different by one point? (Report your answer as a percent.)

24.8%

- d. Interpret the coefficient on the variable $\log(\text{libvol})$.

It's an elasticity.

- e. Would you say it is better to attend a higher ranked law school? How much is a difference in ranking of 20 worth in terms of predicted starting salary?

It's better to be at a better law school (ranking closer to 1). A 20 point difference in ranking leads to a 20(0.0033)% difference in starting salary.

10. Suppose the average worker productivity at manufacturing firms (avgprod) depends on two factors: average hours of training (avgtrain) and average worker ability (avgabil):

$$\text{avgprod} = \beta_0 + \beta_1\text{avgtrain} + \beta_2\text{avgabil} + \epsilon$$

Assume that this equation satisfies the Gauss-Markov assumptions. If grants have been given to firms whose workers have less than average ability, so that avgtrain and avgabil are negatively correlated, what is the likely bias in $\hat{\beta}_1$ obtained from the simple regression of avgprod on avgtrain ?

See notes on bias given in the multiple regression handout.

11. The following equation represents the effects of tax revenue mix on subsequent employment growth for the population of counties in the United States:

$$\text{growth} = \beta_0 + \beta_1\text{share}_p + \beta_2\text{share}_i + \beta_3\text{share}_s + \text{other factors},$$

where growth is the percentage change in employment from 1980 to 1990, share_p is the share of property taxes in total tax revenue, share_i is the share of income tax revenues, and share_s is the share of sales tax revenues. All of these variables are measured in 1980 dollars. The omitted share, share_f , includes fees and miscellaneous taxes. By definition, the four shares add up to one. Other

factors would include expenditures on education, infrastructure, and so on (all measured in 1980\$s).

- a. Why must we omit one of the tax share variables from the equation?

Perfect collinearity would result.

- b. Give a careful interpretation of β_1 .

This is tricky. It will be the incremental change in employment if property tax shares were to increase by 1%, holding income tax and sales tax shares constant.