Causal Inference - Problem Set 1

Gustavo Baroni

8/10/2020

Question 1

You are hired by the governor to study whether a tax on liquor has decreased average liquor consumption in your state. You are able to obtain, for a sample of individuals selected at random, the difference in liquor consumption (in ounces) for the years before and after the tax. For person i who is sampled randomly from the population, Y_i denotes the change in liquor consumption. Treat these as a random sample from a $Normal(\mu, \sigma^2)$ distribution.

(a) The null hypothesis is that there was no change in average liquor consumption. State this formally in terms of μ.

$H_0 : \mu = 0$

(b) The alternative is that there was a decline in liquor consumption; state the alternative in terms of μ.

$H_1 : \mu < 0$

(c) Now, suppose your sample size is $n = 900$ and you obtain the estimates $\hat{y} = -32.8$ and $s = 466.4$. Calculate the t statistic for testing H_0 against H_1; obtain the p-value for the test. (Because of the large sample size, just use the standard normal distribution tabulated in Table G.1.) Do you reject H_0 at the 5% level? At the 1% level?

I did not know even how to start this exercise, so I watched some informative videos from Khan Academy to learn more about t statistic.

From the best of my knowledge, t statistic is used with small samples and, on the other hand, z statustic is used with large ones. Besides that, Khan Academy suggests that $n > 30$ is already a large sample.

Since the sample size is 900 and the exercise is asking for t statistic, I am not sure whether I have done the best research.

Calculating the t statistic:

$$
\begin{aligned}
t &\approx \frac{\hat{y} - \mu_0}{\frac{s}{\sqrt{n}}} \\
&\approx \frac{(-32.8) - 0}{\frac{466.4}{\sqrt{900}}} \\
&\approx \frac{-32.8}{\frac{466.4}{30}} \\
&\approx -2.10
\end{aligned}
$$

Obtaining the p-value using the standard normal distribution table from University of Arizona:

$P(\hat{y} < -2.10) = 0.01786$

Ergo, I do reject H_0 at the 5% level (0.05 is greater than 0.01786), but I do not at the 1% level (0.01 is not greater than 0.01786).

(d) What has been implicitly assumed in your analysis about other determinants of liquor consumption over the two-year period in order to infer causality from the tax change to liquor consumption?

The hipothesis that tax on liquor could be the only or, at least, the most important factor to decrease its consumption has been implicitly assumed in the previous analyses. It is possible to someone claims that because there is no request for an alternative correlation.

Question 2

Using data from 1988 for houses sold in Andover, Massachusetts, from Kiel and McClain (1995), the following equation relates housing price (price) to the distance from a recently built garbage incinerator (dist):

$$\widehat{log(price)} = 9.4 + 0.312log(dist)$$
$$n = 135$$
$$R^2 = 0.162$$

(a) Interpret the coefficient on $log(dist)$. Is the sign of this estimate what you expect it to be?

The coefficient on $log(dist)$ indicates the distance between a house and a garbage incinerator. As I was expecting, the price of a house increases when the place is far from the garbage: mathematically, the first equation says that when the $log(dist)$ increases in one unity, the $\widehat{log(price)}$ increases one unity times 0.312.

(b) Do you think simple regression provides an unbiased estimator of the ceteris paribus elasticity of price with respect to dist?

A simple regression of a simple equation as $y = \beta_0 + \beta_1 x_1 + u$ is not the best method to avoid bisedness because there is not an alternative variable to explain the elasticity of price with respect to distance.

(c) What other factors about a house affect its price? Might these be correlated with distance from the incinerator?

There might well be many factors that affect a price of a house. Among other things, the criminality index in the neighborhood, the distance between the house and a school, and the distance between the house and the incinerator are plausibles variables.

Question 3

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

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

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 $\beta_5 \leq 0$.

We expect $\beta_5 \leq 0$ because we assume that the salary increases if the rank is low (in terms of number), since a $rank = 1$ means the first place in the rank. Therefore, β_5 might be less than equal to zero to valided the previous logic.

(b) What signs do you expect for the other slope parameters? Justify your answers.

My expectations are the following:

β_0 Posite because I assume that a salary is a positive value
β_1 Posite because I suppose that a high median class grade increases the salary
β_2 Posite because I suppose that a high median grade point average increases the salary
β_3 Posite because I suppose that the access to information contributes to increases the salary
β_4 Posite because I suppose that a high cost to study indicates better qualification and that would increase the salary

---

(c) Using the data in LAWSCH85.RAW, the estimated equation is

$$log(salary) = 8.34 + 0.0047LSAT + 0.248GPA + 0.095log(libvol) + 0.038log(cost) - 0.0033rank + u$$

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

The predicted ceteris paribus difference in salary for schools with a median GPA different by one point is equal to 24.8%.

(d) Interpret the coefficient on the variable $log(libvol)$.

The coefficient on the variable $log(libvol)$ indicates that the difference in salary for schools with a library volume different by one point is equal to 9.5%.

(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?

There is almost no difference in the salary concerning the rank of law schools (it impacts about 0.33).

The difference is about 6.6% from the top ranked.

Question 4

Which of the following can cause OLS estimators to be biased? Which of the following can cause the usual OLS t statistics to be invalid (that is, not to have t distributions under H_0)?

(a) Heteroskedasticity.

Heteroskedasticity means that the "variability of a variable is unequal across the range of values of a second variable that predicts it." Therefore, it can reshape the distribuition.

(b) Omitting an important variable.
(c) A sample correlation coefficient of 0.95 between two independent variables both included in the model.

Question 5

Suppose that the model $pctstck = \beta_0 + \beta_1 funds + \beta_2 risktol + u$ satisfies the first four Gauss-Markov assumptions, where pctstck is the percentage of a worker's pension invested in the stock market, funds is the number of mutual funds that the worker can choose from, and risktol is some measure of risk tolerance (larger risktol means the person has a higher tolerance for risk). If funds and risktol are positively correlated, what is the inconsistency in $\widetilde{\beta_1}$, the slope coefficient in the simple regression of pctstck on funds?

There is a positive inconsistency in $\widetilde{\beta_1}$. It is possible to anyone to argue that because a higher tolerance to risks improves the chances of investiments ($\beta_2 > 0$, therefore). Besides that, since funds and risktol are positively correlated, there is, for sure, a positive bias in the model.

Computer Exercises

You must use R to solve the following questions.

Question 6

Use the data in HPRICE1 to estimate the model

$$price = \beta_0 + \beta_1 sqrft + \beta_2 bdrms + u$$

, where price is the house price measured in thousands of dollars.

```r
## Installing the data
hprice <- read.csv("https://raw.githubusercontent.com/moqri/Smart-Econometrics/master/HPRICE1.csv")
```

(a) Write out the results in equation form.

```r
equation <- lm(price ~ sqrft + bdrms, data = hprice)
## I would really like to use summ(equation), but
## I could not load jtools :(
summary(equation)
```

```
##
## Call:
## lm(formula = price ~ sqrft + bdrms, data = hprice)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -127.627  -42.876   -7.051   32.589  229.003
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -19.31500   31.04662  -0.622    0.536
## sqrft         0.12844    0.01382   9.291 1.39e-14 ***
## bdrms        15.19819    9.48352   1.603    0.113
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 63.04 on 85 degrees of freedom
## Multiple R-squared:  0.6319, Adjusted R-squared:  0.6233
## F-statistic: 72.96 on 2 and 85 DF,  p-value: < 2.2e-16
```

Hence, the results in equation form are the following:

$$y = -19.315 + 0.128 sqrft + 15.198 bdrms + u$$

(b) What is the estimated increase in price for a house with one more bedroom, holding square footage constant?

The estimated increase in price for a house with one more bedroom is about 15.2.

(c) What is the estimated increase in price for a house with an additional bedroom that is 140 square feet in size? Compare this to your answer in part (b).

The estimated increase in price for a house with an additional bedroom that is 140 square feet in size is given by the following calculation:

$$\begin{aligned} y &= -19.315 + 0.128 sqrft + 15.198 bdrms + u \\ y &= -19.315 + 0.128 * 140 + 15.198 * 1 + u \\ y &\approx 13.803 \end{aligned}$$

Ergo, un increase of almost 71.5 times the y value in the intercept (-19.315). Besides that, it is possible to affirm that an increase in the square feet has a positive impact in the price.

(d) What percentage of the variation in price is explained by square footage and number of bedrooms?

The percentage of variation in price that is explained by square footage and number of bedrooms is 63.19%.

(e) The first house in the sample has $sqrft = 2,438$ and $bdrms = 4$. Find the predicted selling price for this house from the OLS regression line.

The predicted selling price for the first house in the sample is equal to 354.61.

(f) The actual selling price of the first house in the sample was \$300,000 (so price = 300). Find the residual for this house. Does it suggest that the buyer underpaid or overpaid for the house?

The residual for the first house is given by the difference between the expected value and the real one. In this case, the residual is equal to \$54.60, and, therefore, the buyer underpaid for the house.

Question 7

Use the data in MLB1 for this exercise.

For some unkowning reason, I am not able to use my tidyverse package. I found a good video teaching how to use foreign as an alternative.

```
library(foreign)
mlb1 <- read.dta("C://Users//baron//Downloads//mlb1.dta")
```

(a) Use the model estimated in equation (4.31) and drop the variable rbisyr. What happens to the statistical significance of hrunsyr? What about the size of the coefficient on hrunsyr?

The model estimated in equation 4.31 is this one:

$$\begin{aligned} \widehat{log(salary)} &= 11.19 + .0689years + .0126gamesyr + .00098bavg + .0144hrunsyr + .108rbisyr \\ n &= 355 \\ SSR &= 183.186 \\ R^2 &= .6278 \end{aligned}$$

Here, it is possible to see the statistical significance of each variable:

```
model1 <- lm(log(salary) ~ years+gamesyr+bavg+hrunsyr+rbisyr, data = mlb1)
summary(model1)
```

```
##
## Call:
## lm(formula = log(salary) ~ years + gamesyr + bavg + hrunsyr +
##     rbisyr, data = mlb1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.02508 -0.45034 -0.04013  0.47014  2.68924
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.119e+01  2.888e-01  38.752  < 2e-16 ***
## years       6.886e-02  1.211e-02   5.684 2.79e-08 ***
## gamesyr     1.255e-02  2.647e-03   4.742 3.09e-06 ***
## bavg        9.786e-04  1.104e-03   0.887    0.376
## hrunsyr     1.443e-02  1.606e-02   0.899    0.369
## rbisyr      1.077e-02  7.175e-03   1.500    0.134
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7266 on 347 degrees of freedom
## Multiple R-squared:  0.6278, Adjusted R-squared:  0.6224
## F-statistic: 117.1 on 5 and 347 DF,  p-value: < 2.2e-16
```

```
model2 <- lm(log(salary) ~ years+gamesyr+bavg+hrunsyr, data = mlb1)
```

Dropping rbisyr, the new model is

$$\widehat{log(salary)} = 11.02 + 0.0677years + 0.0158gamesyr + 0.00142bavg + 0.0359hrunsyr$$

Besides that, it is possible to see below that hrunsyr has a statistic significance and the triple impact that it had when rbisyr was in the equation.

```
summary(model2)
```

```
##
## Call:
## lm(formula = log(salary) ~ years + gamesyr + bavg + hrunsyr,
##     data = mlb1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.0642 -0.4614 -0.0271  0.4654  2.7216
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.020913   0.265719  41.476  < 2e-16 ***
## years        0.067732   0.012113   5.592 4.55e-08 ***
## gamesyr      0.015759   0.001564  10.079  < 2e-16 ***
## bavg         0.001419   0.001066   1.331    0.184
## hrunsyr      0.035943   0.007241   4.964 1.08e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7279 on 348 degrees of freedom
## Multiple R-squared:  0.6254, Adjusted R-squared:  0.6211
## F-statistic: 145.2 on 4 and 348 DF,  p-value: < 2.2e-16
```

(b) Add the variables runsyr (runs per year), fldperc (fielding percentage), and sbasesyr (stolen bases per year) to the model from part (i). Which of these factors are individually significant?

Adding the variables:

```
model3 <- lm(log(salary) ~ years+gamesyr+bavg+hrunsyr+runsyr+fldperc+sbasesyr, data = mlb1)
summary(model3)
```

```
##
## Call:
## lm(formula = log(salary) ~ years + gamesyr + bavg + hrunsyr +
##     runsyr + fldperc + sbasesyr, data = mlb1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.11554 -0.44557 -0.08808  0.48731  2.57872
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.4082677  2.0032546   5.196 3.50e-07 ***
## years        0.0699848  0.0119756   5.844 1.18e-08 ***
## gamesyr      0.0078995  0.0026775   2.950 0.003391 **
```

```
## bavg         0.0005296  0.0011038   0.480 0.631656
## hrunsyr      0.0232106  0.0086392   2.687 0.007566 **
## runsyr       0.0173922  0.0050641   3.434 0.000666 ***
## fldperc      0.0010351  0.0020046   0.516 0.605936
## sbasesyr    -0.0064191  0.0051842  -1.238 0.216479
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7176 on 345 degrees of freedom
## Multiple R-squared:  0.639,  Adjusted R-squared:  0.6317
## F-statistic: 87.25 on 7 and 345 DF,  p-value: < 2.2e-16
```

years, runsyr, gamesyr, and hrunsyr are, therefore, individually significant.

(c) In the model from part (b), test the joint significance of bavg, fldperc, and sbasesyr.

```
summary(model3)
```

```
##
## Call:
## lm(formula = log(salary) ~ years + gamesyr + bavg + hrunsyr +
##     runsyr + fldperc + sbasesyr, data = mlb1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.11554 -0.44557 -0.08808  0.48731  2.57872
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.4082677  2.0032546   5.196 3.50e-07 ***
## years        0.0699848  0.0119756   5.844 1.18e-08 ***
## gamesyr      0.0078995  0.0026775   2.950 0.003391 **
## bavg         0.0005296  0.0011038   0.480 0.631656
## hrunsyr      0.0232106  0.0086392   2.687 0.007566 **
## runsyr       0.0173922  0.0050641   3.434 0.000666 ***
## fldperc      0.0010351  0.0020046   0.516 0.605936
## sbasesyr    -0.0064191  0.0051842  -1.238 0.216479
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7176 on 345 degrees of freedom
## Multiple R-squared:  0.639,  Adjusted R-squared:  0.6317
## F-statistic: 87.25 on 7 and 345 DF,  p-value: < 2.2e-16
```

```
model4 <- lm(log(salary) ~ bavg+fldperc+sbasesyr, data = mlb1)
summary(model4)
```

```
##
## Call:
## lm(formula = log(salary) ~ bavg + fldperc + sbasesyr, data = mlb1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.4127 -0.8504 -0.0440  0.9184  2.3216
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.899057   2.931406   1.330  0.18435
## bavg        0.007144   0.001535   4.655 4.61e-06 ***
## fldperc     0.007648   0.002921   2.618  0.00923 **
## sbasesyr    0.032330   0.005282   6.121 2.50e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.068 on 349 degrees of freedom
## Multiple R-squared:  0.1911, Adjusted R-squared:  0.1841
## F-statistic: 27.48 on 3 and 349 DF,  p-value: 5.596e-16
```

The joint significance of bavg, fldperc, and sbasesyr is given by the following:

$$\begin{aligned} F &= \frac{(R_U^2 - R_R^2)/(df_R - df_U)}{(1 - R_U^2)/(n - k - 1)} \\ &= \frac{(0.639^2 - 0.1911^2)/(349 - 345)}{(1 - 0.639)/(353 - 7 - 1)} \\ &\approx 88.83 \end{aligned}$$