Problem Set 3 - Solutions

Statistics and Econometrics

Question 1

Use the data in kielmc.RData, only for the year 1981, to answer the following questions. The data are for houses that sold during 1981 in North Andover, Massachusetts; 1981 was the year construction began on a local garbage incinerator.

1. To study the effects of the incinerator location on housing price, consider the simple regression model

$$\log(price) = \beta_0 + \beta_1 \log(dist) + u,$$

   where *price* is housing price in dollars and *dist* is distance from the house to the incinerator measured in feet. Interpreting this equation causally, what sign do you expect for β_1 if the presence of the incinerator depresses housing prices? Estimate this equation and interpret the results.
2. To the simple regression model in part 1, add the variables $\log(intst)$, $\log(area)$, $\log(land)$, *rooms*, *baths*, and *age*, where *intst* is distance from house to interstate (i.e., a major system of highways running between US states) entrance ramp measured in feet, *area* is square footage of the house, *land* is the lot size in square feet, *rooms* is total number of rooms, *baths* is number of bathrooms, and *age* is age of the house in years. Now, what do you conclude about the effects of the incinerator? Explain why parts 1 and 2 give conflicting results.
3. Add $[\log(intst)]^2$ to the model from part 2. Now what happens? What do you conclude about the importance of functional form?
4. Is the square of $\log(dist)$ significant when you add it to the model from part 3?

Solutions

1. We would expect that $\beta_1 \geq 0$ if the presence of the incinerator depresses housing prices. The causal effect of *dist* on *price* means that: all other relevant factors equal, if the distance from the house to the incinerator increases by 1%, the housing price is expected to increase by β_1%. In other words, it is better to have a home farther away from the incinerator.

```
load("kielmc.RData")
data.1981 <- data %>% filter(year == 1981)
price.m1 <- lm(log(price) ~ log(dist), data = data.1981)
price.m2 <- lm(log(price) ~ log(dist) + log(intst) + log(area) + log(land) + rooms
               + baths + age, data = data.1981)
price.m3 <- lm(log(price) ~ log(dist) + log(intst) + I(log(intst)^2) + log(area)
               + log(land) + rooms + baths + age, data = data.1981)
price.m4 <- lm(log(price) ~ log(dist) + I(log(dist)^2) + log(intst) + I(log(intst)^2)
               + log(area) + log(land) + rooms + baths + age  , data = data.1981)
stargazer(price.m1, price.m2, price.m3, price.m4, font.size = "small",
          header = FALSE, type = 'latex', title = 'Question 1')
```

The result (Model (1) in Table 1) shows that a 1% increase in distance from the incinerator is associated with a predicted price that is about .37% higher.

Table 1: Question 1

	Dependent variable:			
	log(price)			
	(1)	(2)	(3)	(4)
log(dist)	0.365^{***}	0.055	0.185^{***}	0.870
	(0.066)	(0.058)	(0.062)	(2.071)
I(log(dist)^2)				−0.036
				(0.110)
log(intst)		−0.039	2.073^{***}	1.934^{***}
		(0.052)	(0.501)	(0.654)
I(log(intst)^2)			-0.119^{***}	-0.111^{***}
			(0.028)	(0.038)
log(area)		0.319^{***}	0.359^{***}	0.355^{***}
		(0.076)	(0.073)	(0.074)
log(land)		0.077^{*}	0.091^{**}	0.088^{**}
		(0.040)	(0.037)	(0.039)
rooms		0.043	0.038	0.038
		(0.028)	(0.027)	(0.027)
baths		0.167^{***}	0.150^{***}	0.151^{***}
		(0.042)	(0.040)	(0.040)
age		-0.004^{***}	-0.003^{***}	-0.003^{***}
		(0.001)	(0.001)	(0.001)
Constant	8.047^{***}	7.592^{***}	−3.318	−5.916
	(0.646)	(0.642)	(2.646)	(8.296)
Observations	142	142	142	142
R^2	0.180	0.748	0.778	0.778
Adjusted R^2	0.174	0.734	0.764	0.763
Residual Std. Error	0.354 (df = 140)	0.201 (df = 134)	0.189 (df = 133)	0.190 (df = 132)
F Statistic	30.786^{***} (df = 1; 140)	56.683^{***} (df = 7; 134)	58.111^{***} (df = 8; 133)	51.320^{***} (df = 9; 132)
Note:				*p<0.1; **p<0.05; ***p<0.01

2. The result is given by Model (2) in Table 1. When the variables $\log(inst)$, $\log(area)$, $\log(land)$, $rooms$, $baths$, and age are added to the regression, the coefficient on $\log(dist)$ becomes about .055 (se $\approx$ .058). The effect is much smaller now and is statistically insignificant. This is because we have explicitly controlled for several other factors that determine the quality of a home (such as its size and number of baths) and its location (distance to the interstate). This is consistent with the hypothesis that the incinerator was located near less desirable homes to begin with.

3. The result is given by Model (3) in Table 1. When $[\log(inst)]^2$ is added to the regression in part 2, the coefficient on $\log(dist)$ is now very statistically significant, with a t statistic of about three. The coefficients on $\log(inst)$ and $[\log(inst)]^2$ are both very statistically significant, each with t statistics above four in absolute value. Just adding $[\log(inst)]^2$ has had a very big effect on the coefficient. This means that distance from the incinerator and distance from the interstate are correlated in some nonlinear way that also affects housing price.

We can find the value of $\log(inst)$ where the effect on $\log(price)$ actually becomes negative: $2.073/[2(.1193)] \approx 8.69$. When we exponentiate this we obtain about 5,943 feet from the interstate. Therefore, it is best to have your home away from the interstate just over a mile. After that, moving farther away from the interstate lowers predicted house price.

4. The result is given by Model (4) in Table 1. The coefficient on $[\log(dist)]^2$, when it is added to the model estimated in part 3, is about -.0365, but its t statistic is only about -.33. Therefore, it is not necessary to add this complication.

Question 2

Use the data in gpa2.RData for this exercise. Consider the equation

$$colgpa = \beta_0 + \beta_1 hsize + \beta_2 hsize^2 + \beta_3 hsperc + \beta_4 sat + \beta_5 female + \beta_6 athlete + u,$$

where $colgpa$ is cumulative college grade point average; $hsize$ is size of high school graduating class, in hundreds; $hsperc$ is academic percentile in graduating class; sat is combined SAT score; $female$ is a binary gender variable; and $athlete$ is a binary variable, which is one for student athletes.

1. Estimate the equation and report the results. What is the estimated GPA differential between athletes and nonathletes? Is it statistically significant?
2. Drop sat from the model and reestimate the equation. Now, what is the estimated effect of being an athlete? Discuss why the estimate is different than that obtained in part 1.
3. In the model, allow the effect of being an athlete to differ by gender and test the null hypothesis that there is no difference between women athletes and women nonathletes.
4. Does the effect of sat on $colgpa$ differ by gender? Justify your answer.

Solutions

1.

```
load("gpa2.RData")
gpa.m1 <- lm(colgpa ~ hsize + I(hsize^2) + hsperc + sat + female + athlete, data = data)
gpa.m2 <- lm(colgpa ~ hsize + I(hsize^2) + hsperc + female + athlete, data = data)
gpa.m3 <- lm(colgpa ~ hsize + I(hsize^2) + hsperc + sat + female
             + athlete + female:athlete, data = data)
gpa.m4 <- lm(colgpa ~ hsize + I(hsize^2) + hsperc + sat*female
                        + athlete, data = data)
stargazer(gpa.m1, gpa.m2, gpa.m3, gpa.m4, font.size = "small", omit.stat = "f",
          header = FALSE, type = 'latex', title = 'Question 2')
```

Table 2: Question 2

	Dependent variable:			
	colgpa			
	(1)	(2)	(3)	(4)
hsize	-0.057^{***}	-0.053^{***}	-0.057^{***}	-0.057^{***}
	(0.016)	(0.018)	(0.016)	(0.016)
I(hsizeˆ2)	0.005^{**}	0.005^{**}	0.005^{**}	0.005^{**}
	(0.002)	(0.002)	(0.002)	(0.002)
hsperc	-0.013^{***}	-0.017^{***}	-0.013^{***}	-0.013^{***}
	(0.001)	(0.001)	(0.001)	(0.001)
sat	0.002^{***}		0.002^{***}	0.002^{***}
	(0.0001)		(0.0001)	(0.0001)
female	0.155^{***}	0.058^{***}	0.155^{***}	0.102
	(0.018)	(0.019)	(0.018)	(0.134)
athlete	0.169^{***}	0.005	0.167^{***}	0.168^{***}
	(0.042)	(0.045)	(0.048)	(0.043)
female:athlete			0.008	
			(0.096)	
sat:female				0.0001
				(0.0001)
Constant	1.241^{***}	3.048^{***}	1.242^{***}	1.264^{***}
	(0.079)	(0.033)	(0.080)	(0.097)
Observations	4,137	4,137	4,137	4,137
R^2	0.293	0.189	0.293	0.293
Adjusted R^2	0.291	0.188	0.291	0.291
Residual Std. Error	0.554 (df = 4130)	0.594 (df = 4131)	0.554 (df = 4129)	0.554 (df = 4129)
Note:				*p<0.1; **p<0.05; ***p<0.01

The result is shown in Model (1) in Table 2. Holding other factors fixed, an athlete is predicted to have a GPA about .169 points higher than a nonathlete. The t statistic .169/.042 $\approx$ 4.02, which is very significant.

2. With *sat* dropped from the model (Model (2) in Table 2), the coefficient on *athlete* becomes about .0054 (se $\approx$ .0448), which is practically and statistically not different from zero. We can explain the difference based on omitted variable bias. In Model (1), we know that SAT score result is positively correlated with a student's college GPA. On average, we would expect that athletes score lower than non-athletes in SAT. So when we omit SAT from the model, we would have a negative bias. That is, we underestimate the difference in college performance between athletes and non-athletes. Part 1 shows that, once we account for SAT differences, athletes do better than non-athletes.

3. To facilitate testing the hypothesis that there is no difference between women athletes and women nonathletes, we include $female$, *athlete* and the interaction term between $female$ and *athlete* in the model. The result is shown in Model (3) in Table 2. The null hypothesis that there is no difference between female athletes and female nonathletes can be written as $H_0 : \beta_{athlete} + \beta_{female \cdot athlete} = 0$. We run the test using the linearHypothesis() function.

```
linearHypothesis(gpa.m3, c("athlete + female:athlete = 0"))
```

```
## Linear hypothesis test
##
## Hypothesis:
## athlete  + female:athlete = 0
##
## Model 1: restricted model
## Model 2: colgpa ~ hsize + I(hsize^2) + hsperc + sat + female + athlete +
##     female:athlete
##
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1   4130 1270.7
## 2   4129 1269.4  1    1.3352 4.3431 0.03722 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can reject the null hypothesis at 5% significance level, which means that there is (weak) evidence for the difference in college GPA between female atheles and female nonathletes.

4. Whether we add the interaction $female \cdot sat$ to the equation in part 1 or part 3, the outcome is practically the same. For example, when $female \cdot sat$ is added to the equation in part 1 (Model (4) in Table 2), its coefficient is about .000051 and its t statistic is about .40. Thus, there is very little evidence that the effect of *sat* differs by gender.

Question 3

Consider the following model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u.$$

Our goal is to understand the causal impact of x_1 on y. However, there is an issue of multicollinearity in the model, meaning that x_1 and x_2 are highly correlated (say, correlation between the two is greater than 0.95), and thus we cannot correctly estimate the partial impact of x_1 on y due to inflated standard errors. Discuss how we can estimate the causal impact of x_1 on y in this situation.

Solutions

We can potentially drop x_2, and simply regress y on x_1, where the impact of x_1 on y is reflected by its coefficient. The reason why omitted variable bias is not too big of a concern here is because when the

correlation between two variables are as high as 0.95, it typically means that they are measuring the same thing. So after controlling for x_1, x_2 becomes irrelevant, meaning that its partial impact on y is very close to 0. Thus, based on the formula for omitted variable bias, the bias $\tilde{\delta}\beta_2$ from omitting x_2 is very small.