HW10_Sampathirao_A

Anvita Sampathirao

7/30/2019

#1.1

```
library("readxl")
colgdata<- read_excel("CollegeDistance.xls", col_names = TRUE)
suppressMessages(attach(colgdata))
suppressMessages(library(stargazer))
Reg1 <- lm(ed ~ dist + female + bytest + tuition + black + hispanic + incomehi + ownhome + dadcoll + mo
stargazer(Reg1,
          type= "latex",
          intercept.bottom = FALSE)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Tue, Jul 30, 2019 - 11:50:05 PM

Because dist is a linear regressor on ed, a unit change in dist (be it from 2 to 3 or 6 to 7) will result in a decrease in ed by a factor of 0.037.

i.e.

$$\Delta y = \beta_1 * \Delta x_1$$

Now when delta_x = 1, delta_y = beta_1

where beta_1 is the estimated coefficient of dist which is -0.037 in our case.

#1.2

```
Reg2 <- lm(ed ~ dist + I(dist^2) + female + bytest + tuition + black + hispanic + incomehi + ownhome + 
stargazer(Reg2,
          type= "latex",
          intercept.bottom = FALSE)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Tue, Jul 30, 2019 - 11:50:05 PM

Because dist is a quadratic regressor on ed, a unit change in dist will result in change in ed by -0.081 + (2 * 0.005 * dist)

i.e.

$$\Delta y = (\beta_1 + 2\beta_2 x_1) * \Delta x_1$$

Where, x1 is dist and

beta_1 is estimated coefficient of dist (linear regressor) = -0.081

beta_2 is estimated coefficient of dist (quadratic regressor) = 0.005

Now when dist is changing from 2 to 3,

delta_x = 1

Table 1:

	Dependent variable:
	ed
Constant	8.921^{***}
	(0.252)
dist	-0.037^{***}
	(0.013)
female	0.143^{***}
	(0.050)
bytest	0.093^{***}
	(0.003)
tuition	-0.191^{*}
	(0.101)
black	0.351^{***}
	(0.071)
hispanic	0.362^{***}
	(0.077)
incomehi	0.372^{***}
	(0.061)
ownhome	0.139^{**}
	(0.067)
dadcoll	0.571^{***}
	(0.074)
momcoll	0.378^{***}
	(0.082)
cue80	0.029^{***}
	(0.010)
stwmfg80	-0.043^{**}
	(0.020)
Observations	3,796
R^2	0.284
Adjusted R^2	0.281
Residual Std. Error	1.538 (df = 3783)
F Statistic	124.806^{***} (df = 12; 3783)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 2:

	Dependent variable:
	ed
Constant	9.012^{***}
	(0.256)
dist	-0.081^{***}
	(0.026)
I(distˆ2)	0.005^{**}
	(0.002)
female	0.143^{***}
	(0.050)
bytest	0.093^{***}
	(0.003)
tuition	-0.193^{*}
	(0.101)
black	0.334^{***}
	(0.072)
hispanic	0.333^{***}
	(0.079)
incomehi	0.369^{***}
	(0.061)
ownhome	0.143^{**}
	(0.067)
dadcoll	0.561^{***}
	(0.074)
momcoll	0.378^{***}
	(0.081)
cue80	0.026^{***}
	(0.010)
stwmfg80	-0.043^{**}
	(0.020)
Observations	3,796
R^2	0.284
Adjusted R^2	0.282
Residual Std. Error	1.537 (df = 3782)
F Statistic	115.608^{***} (df = 13; 3782)
Note:	*p<0.1; **p<0.05; ***p<0.01

x_1 = 2

Therefore,

$$\Delta y = (-0.081 + 2 * 0.005 * 2) * 1 = -0.061$$

Now when dist is changing from 6 to 7,

delta_x = 1

x_1 = 6

Therefore,

$$\Delta y = (-0.081 + 2 * 0.005 * 6) * 1 = -0.021$$

As compared to (1), in model (2), as dist increases, the rate of change of ed is decreasing with a change in dist.

If the relation between ed and dist is linear, our hypothesis will be:

$$H_0 : \beta_2 = 0$$

which from the above regression results we observe that,

$$\beta_2 = 0.005^{**}$$

It implies we can reject our null hypothesis at a 95% significance level and thus, the relation between ed and dist is non-linear.

#1.3

```
Reg3 <- lm(ed ~ dist + I(dist^2) + female + bytest + tuition + black + hispanic + incomehi + ownhome + d
stargazer(Reg3,
          type= "latex",
          intercept.bottom = FALSE)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Tue, Jul 30, 2019 - 11:50:05 PM

We see that,

$$\beta_{dadcoll} = 0.654^{***}$$

That means, when father is educated, years of college education increase by 65.4% as compared to when father is not educated and is statistically significant at 99%

Also,

$$\beta_{momcoll} = 0.569^{***}$$

That means, when mother is educated, years of college education increase by 56.9% as compared to when mother is not educated and is statistically significant at 99%

Now, when we add an interaction term between dadcoll and momcoll,

$$\beta_{momcoll*dadcoll} = -0.366^{**}$$

Table 3:

	Dependent variable:
	ed
Constant	9.002^{***}
	(0.255)
dist	-0.081^{***}
	(0.026)
I(distˆ2)	0.005^{**}
	(0.002)
female	0.141^{***}
	(0.050)
bytest	0.093^{***}
	(0.003)
tuition	-0.194^{*}
	(0.101)
black	0.331^{***}
	(0.072)
hispanic	0.330^{***}
	(0.079)
incomehi	0.362^{***}
	(0.061)
ownhome	0.141^{**}
	(0.067)
dadcoll	0.654^{***}
	(0.084)
momcoll	0.569^{***}
	(0.117)
I(momcoll ∗dadcoll)	-0.366^{**}
	(0.161)
cue80	0.026^{***}
	(0.010)
stwmfg80	-0.042^{**}
	(0.020)
Observations	3,796
R^2	0.285
Adjusted R^2	0.283
Residual Std. Error	1.536 (df = 3781)
F Statistic	107.837^{***} (df = 14; 3781)
Note:	*p≤0.1; **p<0.05; ***p<0.01

#Not statistically significant at 99%

That means, when mother and father both are educated, years of college education increase by

$$\beta_{momcoll} + \beta_{dadcoll} + \beta_{momcoll*dadcoll} = 0.654 + 0.569 - 0.366 = 0.857$$

85.7% as compared to when mother and father both are not educated.

#1.4

```
Reg4 <- lm(ed ~ dist + I(dist^2) + female + bytest + tuition + black + hispanic + incomehi + I(dist*inc
stargazer(Reg4,
          type= "latex",
          intercept.bottom = FALSE)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Tue, Jul 30, 2019 - 11:50:05 PM

The interaction parameter is not statistically significant at 95%.

#1.5

```
anova(Reg2, Reg3)
```

Analysis of Variance Table

Model 1: ed ~ dist + I(dist^2) + female + bytest + tuition + black + hispanic + incomehi + ownhome + dadcoll + momcoll + cue80 + stwmfg80 Model 2: ed ~ dist + I(dist^2) + female + bytest + tuition + black + hispanic + incomehi + ownhome + dadcoll + momcoll + I(momcoll * dadcoll) + cue80 + stwmfg80 Res.Df RSS Df Sum of Sq F Pr(>F)
1 3782 8936.3
2 3781 8924.1 1 12.177 5.1593 0.02318 * — Signif. codes: 0 ‘’ ***0.001*** ’’ *0.01* ” 0.05 ‘.’ 0.1 ’ ’ 1

```
stargazer(anova(Reg2, Reg3), summary = FALSE, type = "latex", header = FALSE, title = "Anova 1")
```

For model (2) nested specification of the complete model (3) we reject the null that, jointly, the parameters of the complete model that are not in the nested model are equal to zero.

That is, the interaction term between momcoll and dadcoll should be included in the model.

#1.6

```
anova(Reg2, Reg4)
```

Analysis of Variance Table

Model 1: ed ~ dist + I(dist^2) + female + bytest + tuition + black + hispanic + incomehi + ownhome + dadcoll + momcoll + cue80 + stwmfg80 Model 2: ed ~ dist + I(dist^2) + female + bytest + tuition + black + hispanic + incomehi + I(dist * incomehi) + ownhome + dadcoll + momcoll + cue80 + stwmfg80 Res.Df RSS Df Sum of Sq F Pr(>F)
1 3782 8936.3
2 3781 8928.3 1 7.9295 3.358 0.06696 . — Signif. codes: 0 ‘’ ***0.001*** ’’ *0.01* ” 0.05 ‘.’ 0.1 ’ ’ 1

```
stargazer(anova(Reg2, Reg4), summary = FALSE, type = "latex", header = FALSE, title = "Anova 2")
```

Table 4:

	Dependent variable:
	ed
Constant	9.041***
	(0.256)
dist	−0.100***
	(0.027)
I(distˆ2)	0.005**
	(0.002)
female	0.144***
	(0.050)
bytest	0.093***
	(0.003)
tuition	−0.204**
	(0.101)
black	0.335***
	(0.072)
hispanic	0.330***
	(0.079)
incomehi	0.280***
	(0.078)
I(dist ∗incomehi)	0.057*
	(0.031)
ownhome	0.145**
	(0.067)
dadcoll	0.570***
	(0.074)
momcoll	0.380***
	(0.081)
cue80	0.026***
	(0.010)
stwmfg80	−0.043**
	(0.020)
Observations	3,796
R^2	0.285
Adjusted R^2	0.282
Residual Std. Error	1.537 (df = 3781)
F Statistic	107.657*** (df = 14; 3781)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 5: Anova 1

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	3,782	8,936.257				
2	3,781	8,924.080	1	12.177	5.159	0.023

Table 6: Anova 2

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	3,782	8,936.257				
2	3,781	8,928.327	1	7.930	3.358	0.067

For model (2) nested specification of the complete model (4) we fail to reject the null that, jointly, the parameters of the complete model that are not in the nested model are equal to zero.

That is, the interaction term between dist and incomehi can be excluded from the model.

#1.7

```
Reg7_a <- lm(ed ~ dist + I(dist^2) + I(dist^3) + female + bytest + tuition + black + hispanic + incomehi
Reg7_b <- lm(ed ~ dist + I(dist^2) + I(dist^3) + I(dist^4) + female + bytest + tuition + black + hispani
Regs<- list (Reg1, Reg2, Reg7_a, Reg7_b)
stargazer(Regs,
          type= "latex",
          df= FALSE,
          intercept.bottom = FALSE)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Tue, Jul 30, 2019 - 11:50:05 PM

We see from above that, beta_1,beta_2,beta_3,beta_4 are insignificant in models (3) & (4)

where, beta_1= estimated coeff of dist

beta_2= estimated coeff of distˆ2

beta_3= estimated coeff of distˆ3

beta_4= estimated coeff of distˆ4

Therefore, relationship between dist and ed cannot be greater than order 2 in our case.

#2.8

```
a<- .Machine$double.xmin
Reg8 <- lm(ed ~ I(log(dist+a)) + female + bytest + tuition + black + hispanic + incomehi + ownhome + da
stargazer(Reg8,
          type= "latex",
          intercept.bottom = FALSE)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Tue, Jul 30, 2019 - 11:50:05 PM

#2.9

Table 7:

	Dependent variable:			
	ed			
	(1)	(2)	(3)	(4)
Constant	8.921***	9.012***	9.001***	9.010***
	(0.252)	(0.256)	(0.257)	(0.258)
dist	−0.037***	−0.081***	−0.058	−0.085
	(0.013)	(0.026)	(0.053)	(0.089)
I(dist^2)		0.005**	−0.001	0.011
		(0.002)	(0.012)	(0.034)
I(dist^3)			0.0003	−0.001
			(0.001)	(0.004)
I(dist^4)				0.0001
				(0.0002)
female	0.143***	0.143***	0.143***	0.143***
	(0.050)	(0.050)	(0.050)	(0.050)
bytest	0.093***	0.093***	0.093***	0.093***
	(0.003)	(0.003)	(0.003)	(0.003)
tuition	−0.191*	−0.193*	−0.191*	−0.188*
	(0.101)	(0.101)	(0.101)	(0.101)
black	0.351***	0.334***	0.337***	0.335***
	(0.071)	(0.072)	(0.072)	(0.072)
hispanic	0.362***	0.333***	0.333***	0.331***
	(0.077)	(0.079)	(0.079)	(0.079)
incomehi	0.372***	0.369***	0.370***	0.370***
	(0.061)	(0.061)	(0.061)	(0.061)
ownhome	0.139**	0.143**	0.143**	0.143**
	(0.067)	(0.067)	(0.067)	(0.067)
dadcoll	0.571***	0.561***	0.561***	0.561***
	(0.074)	(0.074)	(0.074)	(0.074)
momcoll	0.378***	0.378***	0.377***	0.377***
	(0.082)	(0.081)	(0.082)	(0.082)
cue80	0.029***	0.026***	0.026**	0.026**
	(0.010)	(0.010)	(0.010)	(0.010)
stwmfg80	−0.043**	−0.043**	−0.043**	−0.043**
	(0.020)	(0.020)	(0.020)	(0.020)
Observations	3,796	3,796	3,796	3,796
R^2	0.284	9 0.284	0.284	0.284
Adjusted R^2	0.281	0.282	0.282	0.282
Residual Std. Error	1.538	1.537	1.537	1.537
F Statistic	124.806***	115.608***	107.347***	100.177***

Table 8:

	Dependent variable:
	ed
Constant	8.800^{***}
	(0.248)
I(log(dist + a))	-0.00000
	(0.0003)
female	0.146^{***}
	(0.050)
bytest	0.093^{***}
	(0.003)
tuition	-0.120
	(0.098)
black	0.378^{***}
	(0.071)
hispanic	0.387^{***}
	(0.077)
incomehi	0.377^{***}
	(0.061)
ownhome	0.131^{**}
	(0.067)
dadcoll	0.583^{***}
	(0.074)
momcoll	0.381^{***}
	(0.082)
cue80	0.020^{**}
	(0.009)
stwmfg80	-0.040^{*}
	(0.020)
Observations	3,796
R^2	0.282
Adjusted R^2	0.280
Residual Std. Error	1.539 (df = 3783)
F Statistic	123.841^{***} (df = 12; 3783)
Note:	*p<0.1; **p<0.05; ***p<0.01

```
Reg9 <- lm(I(log(ed)) ~ dist + female + bytest + tuition + black + hispanic + incomehi + ownhome + dadc
stargazer(Reg9,
          type= "latex",
          intercept.bottom = FALSE)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Tue, Jul 30, 2019 - 11:50:06 PM

#2.10

```
Reg10 <- lm(I(log(ed)) ~ I(log(dist+a)) + female + bytest + tuition + black + hispanic + incomehi + ownl
stargazer(Reg10,
          type= "latex",
          intercept.bottom = FALSE)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Tue, Jul 30, 2019 - 11:50:06 PM

#2.11

```
Regs_1<- list(Reg8, Reg9, Reg10)
stargazer(Regs_1,
          type= "latex",
          df= FALSE,
          intercept.bottom = FALSE)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Tue, Jul 30, 2019 - 11:50:06 PM

Model (1): Linear-Log Model: Though the sign suggests a decreasing nature, a unit increase in dist will have a negligent reduction in the years of college education. Almost no change.

Model (2): Log-Linear Model: A unit increase in dist, will have a reduction by a factor of 0.3% in years of college eduction. Decreasing nature.

Model (3): Log-Log Model: A unit increase in dist has no change in years of college education. Parameter is 0 and also statistically insignificant.

#2.12

```
SSE1<- sum(ed- Reg1$fitted.values)^2
SSE2<- sum(ed- Reg2$fitted.values)^2
SSE3<- sum(ed- Reg3$fitted.values)^2
SSE4<- sum(ed- Reg4$fitted.values)^2
SSE8<- sum(ed- Reg8$fitted.values)^2
SSE9<- sum(ed- (exp(Reg9$fitted.values)))^2
SSE10<- sum(ed- (exp(Reg10$fitted.values)))^2

SSEdisplay<- c(SSE1, SSE2, SSE3, SSE4, SSE8, SSE9, SSE10)
SSEdisplay
```

```
## [1] 1.878625e-25 4.029186e-26 9.895471e-27 7.005400e-26 1.135960e-26
## [6] 9.928670e+04 9.971758e+04
```

Table 9:

	Dependent variable:
	I(log(ed))
Constant	2.266^{***}
	(0.018)
dist	-0.003^{***}
	(0.001)
female	0.010^{***}
	(0.004)
bytest	0.007^{***}
	(0.0002)
tuition	-0.014^{*}
	(0.007)
black	0.026^{***}
	(0.005)
hispanic	0.026^{***}
	(0.005)
incomehi	0.027^{***}
	(0.004)
ownhome	0.010^{**}
	(0.005)
dadcoll	0.041^{***}
	(0.005)
momcoll	0.027^{***}
	(0.006)
cue80	0.002^{***}
	(0.001)
stwmfg80	-0.003^{**}
	(0.001)
Observations	3,796
R^2	0.285
Adjusted R^2	0.283
Residual Std. Error	0.109 (df = 3783)
F Statistic	125.852^{***} (df = 12; 3783)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 10:

	Dependent variable:
	I(log(ed))
Constant	2.257^{***}
	(0.018)
I(log(dist + a))	0.00000
	(0.00002)
female	0.011^{***}
	(0.004)
bytest	0.007^{***}
	(0.0002)
tuition	-0.009
	(0.007)
black	0.028^{***}
	(0.005)
hispanic	0.028^{***}
	(0.005)
incomehi	0.027^{***}
	(0.004)
ownhome	0.009^{*}
	(0.005)
dadcoll	0.041^{***}
	(0.005)
momcoll	0.027^{***}
	(0.006)
cue80	0.001^{**}
	(0.001)
stwmfg80	-0.003^{*}
	(0.001)
Observations	3,796
R^2	0.284
Adjusted R^2	0.281
Residual Std. Error	0.109 (df = 3783)
F Statistic	124.882^{***} (df = 12; 3783)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 11:

	Dependent variable:		
	ed	I(log(ed))	
	(1)	(2)	(3)
Constant	8.800***	2.266***	2.257***
	(0.248)	(0.018)	(0.018)
I(log(dist + a))	−0.00000		0.00000
	(0.0003)		(0.00002)
dist		−0.003***	
		(0.001)	
female	0.146***	0.010***	0.011***
	(0.050)	(0.004)	(0.004)
bytest	0.093***	0.007***	0.007***
	(0.003)	(0.0002)	(0.0002)
tuition	−0.120	−0.014*	−0.009
	(0.098)	(0.007)	(0.007)
black	0.378***	0.026***	0.028***
	(0.071)	(0.005)	(0.005)
hispanic	0.387***	0.026***	0.028***
	(0.077)	(0.005)	(0.005)
incomehi	0.377***	0.027***	0.027***
	(0.061)	(0.004)	(0.004)
ownhome	0.131**	0.010**	0.009*
	(0.067)	(0.005)	(0.005)
dadcoll	0.583***	0.041***	0.041***
	(0.074)	(0.005)	(0.005)
momcoll	0.381***	0.027***	0.027***
	(0.082)	(0.006)	(0.006)
cue80	0.020**	0.002***	0.001**
	(0.009)	(0.001)	(0.001)
stwmfg80	−0.040*	−0.003**	−0.003*
	(0.020)	(0.001)	(0.001)
Observations	3,796	3,796	3,796
R^2	0.282	0.285	0.284
Adjusted R^2	0.280	0.283	0.281
Residual Std. Error	1.539	0.109	0.109
F Statistic	123.841***	125.852***	124.882***
Note:			*p<0.1; **p<0.05; ***p<0.01

```r
min(SSEdisplay)
```

```
## [1] 9.895471e-27
```

Model (3) or Regression (3) has the least SSE, i.e., including the interaction term between momcoll and dadcoll minimizes the SSE.