Problem Set 4

Spencer Papay - ssp2170

2/19/2017

Question 1

a) No, because $sleep + study + work + leisure = 168$, in order for one variable to change, such as *study*, we would have to allow another variable to change as well.

b) This violates the No Perfect Collinearity assumption, specifically the part that states that none of the independent variables are functions of another. Because, for example, with *study*, we can write that $study = 168 - sleep - work - leisure$, and the same is true for any other variable, so this assumption is violated.

c) Remove one of the independent variables, such as *sleep*. Now, β_1 tells us what the effect on GPA is of substituting one more hour of study for one less hour of leisure, when the rest of the variables ($work, leisure$, and u) are fixed. This can be interpreted in the sense that if we hold $leisure$ and $work$ constant, but increasing *study* by an hour, then it must be true that *sleep* reduces by an hour as well. The same holds true for the other variables.

Question 2

a) The dependent variable is one's cognitive performance and growth, including academic achievement during adolescence.

b) The researchers are measuring the casual effect of the frequency of eating fast food meals versus the frequency of eating slow meals per week on socioeconomic status and cognitive performance. They are seeking test the null hypothesis that the main meal of children's days has no significant effect on their cognitive ability/growth. The authors hypothesized that a higher frequency of slow meals per week would be associated with a higher socioeconomic status and better cognitive performance, and on growth between ages 3 and 5.

c) The authors controlled that all participants got the same number of meals per week and completed the same cognitive tests.

d) $Performance_i = \beta_0 + \beta_1 SlowMealFrequency_i + u_i$

e) Socio-economic status was omitted from the example regression above. The author stated that her findings about frequency of fast/slow meals was perhaps a factor of subjects' SES.

f) This variable would be positively correlated with part (b), because as SES goes up, one's family can afford more, or have more time to prepare, slow meals. Beyond part (b), studies suggest that SES actually plays a role in cognitive growth and performance because just as families have more time to prepare slow meals, they have more time to help their children study and complete their homework as well.

g) There is a positive correlation between SES and slow meal frequency ($corr(X_1, X_2) > 0$), and there is a positive correlation between SES and performance ($corr(Y, X_2) > 0$), which means that $corr(X_2, u) > 0$, so the additional variable is overestimated.

Question 3

```
m1 <- lm(ed~dist, data=college)
m2 <- lm(ed~dist+female+black+hispanic, data=college)
m3 <- lm(ed~dist+female+black+hispanic+bytest+incomehi+ownhome+momcoll+dadcoll+cue80+stwmfg80,
         data=college)

cov1 <- vcovHC(m1, type="HC1")
cov2 <- vcovHC(m2, type="HC1")
cov3 <- vcovHC(m3, type="HC1")

se1 <- sqrt(diag(cov1))
se2 <- sqrt(diag(cov2))
se3 <- sqrt(diag(cov3))
```

Table:

```
stargazer(m1, m2, m3, type="latex", header=FALSE, title="Distance to Nearest College",
dep.var.caption = "Education Completed",
omit.stat = c("f","ser"),
se=list(se1,se2,se3))
```

See table on following page for answers to a), b), c).

a) The estimated slope is -0.073.

b) The estimated slope is -0.083.

c) The estimated slope is -0.0308. This coefficient is less than half the magnitude as our original in (a).

d)

```
htest <- linearHypothesis(m3, c("female=0","black=0","hispanic=0","bytest=0",
                                "incomehi=0","ownhome=0",
                                "momcoll=0","dadcoll=0","cue80=0","stwmfg80=0"),
                          vcov=hccm(m3,type="hc1"),test="F")
htest
```

```
## Linear hypothesis test
##
## Hypothesis:
## female = 0
## black = 0
## hispanic = 0
## bytest = 0
## incomehi = 0
## ownhome = 0
## momcoll = 0
## dadcoll = 0
## cue80 = 0
## stwmfg80 = 0
##
## Model 1: restricted model
## Model 2: ed ~ dist + female + black + hispanic + bytest + incomehi + ownhome +
##     momcoll + dadcoll + cue80 + stwmfg80
##
## Note: Coefficient covariance matrix supplied.
```

Table 1: Distance to Nearest College

	Education Completed		
	ed		
	(1)	(2)	(3)
dist	-0.073^{***}	-0.083^{***}	-0.031^{***}
	(0.013)	(0.014)	(0.012)
female		0.006	0.143^{***}
		(0.059)	(0.050)
black		-0.561^{***}	0.354^{***}
		(0.071)	(0.067)
hispanic		-0.205^{**}	0.402^{***}
		(0.085)	(0.074)
bytest			0.092^{***}
			(0.003)
incomehi			0.367^{***}
			(0.062)
ownhome			0.146^{**}
			(0.065)
momcoll			0.379^{***}
			(0.084)
dadcoll			0.570^{***}
			(0.076)
cue80			0.024^{***}
			(0.009)
stwmfg80			-0.050^{**}
			(0.020)
Constant	13.956^{***}	14.108^{***}	8.861^{***}
	(0.038)	(0.055)	(0.241)
Observations	3,796	3,796	3,796
R^2	0.007	0.022	0.283
Adjusted R^2	0.007	0.021	0.281
Note:			*p<0.1; **p<0.05; ***p<0.01

```
##
##   Res.Df Df      F    Pr(>F)
## 1   3794
## 2   3784 10 198.37 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We reject the notion that the coefficients of the variables in (3) are jointly zero. In other words, we reject the notion that other controls do not have an influence.

e) Yes, it appears to suffer from omitted variable bias, because the inclusion of additional variables leads to a significant change in our estimate of β_{dist} while also increasing the magnitude of our R^2.

f) Adjusted R^2, known as $\bar{R}^2$, decreases for each variable that doesn't improve the regression, but its formula $\frac{n-1}{n-k-1}$ outweighs the number of our variables (k=12) by the number of observations we have (n=3796), making it similar to R^2, but less than it. The only time they will be equal is with one variable.

g) The specification in (3) fits the data best. 28% of the variation in the data can be explained due to its included variables, compared to about 2% in the first two specifications. Almost all of the additional controls are additionally significant at 1% p-value.

h) An individual whose father attended college is predicted to attend an additional 0.57 years of education compared to an individual whose father did not attend college.

i) *cue80* represents the level of unemployment in the county, and it is a positive coefficient of 0.02, meaning for a 1 percentage point increase in the county unemployment rate, subjects are estimated to obtain an additional 0.02 years of schooling. *swmfg80* represents the state hourly minimum wage for manufacturing, and is negative, meaning for a \$1/hour increase in minimum manufacturing wages implies subjects are estimated to obtain 0.05 years less of schooling. These effects are relatively small — a \$10/hr increase would be needed to explain a half year less of schooling. Such large changes may be beyond the scope of our regression, though. The directionality of these variables is expected, though: higher immediate wages may convince some students to complete less college, and higher unemployment may lead to a higher demand for education to compete in the skilled labor market.

j)

```
bob <- data.frame(dist=2, female=0, black=1, hispanic=0,
                  bytest=58, incomehi=1, ownhome=1, momcoll=1,
                  dadcoll=0, cue80=7.5, stwmfg80=10)
predict(m3, bob)
```

```
##        1
## 15.08803
```

Bob's predicted years of schooling are 15.09.

k)

```
bob.moved <- bob
bob.moved$dist <- bob.moved$dist+1.5
predict(m3, bob.moved)
```

```
##        1
## 15.04183
```

Bob's predicted years of schooling decreases to 15.04.