HW09_Sampathirao_A

Anvita Sampathirao

7/22/2019

1.1

$$H_0 : \beta_{\text{dist}} = 0$$
$$H_1 : \beta_{\text{dist}} \neq 0$$

```r
library("readxl")
colgdata<- read_excel("CollegeDistance.xls", col_names = TRUE)
#install.packages("stargazer", repos= "http://cran.us.r-project.org")
library("stargazer")
```

```
##
## Please cite as:

##  Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables.

##  R package version 5.2.2. https://CRAN.R-project.org/package=stargazer
```

```r
suppressMessages(attach(colgdata))
##  Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables.
##  R package version 5.2.2. https://CRAN.R-project.org/package=stargazer
Reg1<- lm(ed~dist, data = colgdata)
stargazer(Reg1,
          title= "Bivariate Regression Results",
          dep.var.labels = c("Years of Education Attained"),
          covariate.labels = c("Intercept","Distance from College (in 10's of miles)"),
          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 23, 2019 - 23:05:02

There is a statistically different than 0 relation between distance from college and Years of education attained. If the distance increases by a unit, years of education attained decreases by 0.073 Distance from college is not a very good predictor of years of education attained because R^2 value is very close to 0.

1.2

A regression suffers from an Omitted Variable Bias when the following conditions hold: 1) The Omitted variable is correlated with the included regressor 2) The Omitted variable is a determinant of the dependant variable

Table 1: Bivariate Regression Results

	Dependent variable:
	Years of Education Attained
Intercept	13.956***
	(0.038)
Distance from College (in 10's of miles)	−0.073***
	(0.014)
Observations	3,796
R^2	0.007
Adjusted R^2	0.007
Residual Std. Error	1.807 (df = 3794)
F Statistic	28.476*** (df = 1; 3794)
Note:	*p<0.1; **p<0.05; ***p<0.01

```r
corData <- cor(colgdata, use = "pairwise.complete.obs")
corData <- corData[, colnames(corData) %in% c("ed", "dist")]
colnames(corData) <- c("Distance from College","Years of Education Attained")
row.names(corData) <- c("Gender (0=Male, 1=Female)",
                        "Race (0=Non Black, 1=Black)",
                        "Ethnicity (0= Non Hispanic, 1=Hispanic)",
                        "Test Score",
                        "Father Education (0=Not a College Grad, 1=College Grad)",
                        "Mother Education (0=Not a College Grad, 1=College Grad)",
                        "Family Ownership (0=Do not Own Home, 1=Own Home)",
                        "Schooling (0=Not in Urban area, 1=In Urban area)",
                        "County Unemoloyment Rate in 1980",
                        "State Hourly Wage in Manufacturing in 1980",
                        "Distance from College in 4 years",
                        "Avg State College Tuition in 4 years",
                        "Years of Education Attained",
                        "Family Income (0=Income<=$25000/year, 1=Income>$25000/year)")
stargazer(corData, summary = FALSE,
          type = "latex",
          title = "Correlation Table",
          digits = 2)
```

% 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 23, 2019 - 23:05:03

The bivariate model seems to suffer from an Omitted varibale bias, looking at the correlation between Distance from College and Schooling, County Unemployment Rate, and Schooling. Test scores relatively has a heavy correlation with the dependent variable (Years of Education attained)

1.3

From the correlation table above,

Table 2: Correlation Table

	Distance from College	Years of Education Attai
Gender (0=Male, 1=Female)	-0.003	-0.002
Race (0=Non Black, 1=Black)	-0.10	-0.10
Ethnicity (0= Non Hispanic, 1=Hispanic)	0.02	-0.02
Test Score	-0.06	0.48
Father Education (0=Not a College Grad, 1=College Grad)	-0.11	0.29
Mother Education (0=Not a College Grad, 1=College Grad)	-0.08	0.23
Family Ownership (0=Do not Own Home, 1=Own Home)	0.05	0.09
Schooling (0=Not in Urban area, 1=In Urban area)	-0.30	-0.02
County Unemoloyment Rate in 1980	0.25	-0.01
State Hourly Wage in Manufacturing in 1980	-0.01	0.02
Distance from College in 4 years	1	-0.09
Avg State College Tuition in 4 years	-0.19	0.06
Years of Education Attained	-0.09	1
Family Income (0=Income<=$25000/year, 1=Income>$25000/year)	-0.08	0.22

1) bytest has a strong correlation with our dependent variable ed and simultaneously has a negative correlation with dist
2) urban has a strong correlation with dist and has a negative correlation with ed
3) dadcoll shares a strong correlation with ed and dist
4) The county unemployment rate is strongly correlated with dist

Therefore, bytest, urban, dadcoll and cue80 should be included in the regression.

1.4

```
Reg2<- lm(ed~dist + bytest, data = colgdata)
Reg3<- lm(ed~dist + bytest + urban, data = colgdata)
Reg4<- lm(ed~dist + bytest + urban + dadcoll, data = colgdata)
Reg5<- lm(ed~dist + bytest + urban + dadcoll + cue80, data = colgdata)
Regs<- list(Reg1, Reg2, Reg3, Reg4, Reg5)
stargazer(Regs,
          title = "Regression Results",
          dep.var.labels = c("Years of Eduction Attained"),
          covariate.labels = c("Intercept",
                               "Distance from College",
                               "Test Score",
                               "Schooling",
                               "Father Education",
                               "County Unemployment Rate"),
          type = "latex",
          intercept.bottom = FALSE,
          df= 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 23, 2019 - 23:05:03

Table 3: Regression Results

	Dependent variable:				
	Years of Eduction Attained				
	(1)	(2)	(3)	(4)	(5)
Intercept	13.956***	8.957***	8.916***	9.158***	9.052***
	(0.038)	(0.155)	(0.160)	(0.158)	(0.170)
Distance from College	−0.073***	−0.049***	−0.045***	−0.026**	−0.031**
	(0.014)	(0.012)	(0.013)	(0.013)	(0.013)
Test Score		0.097***	0.098***	0.089***	0.089***
		(0.003)	(0.003)	(0.003)	(0.003)
Schooling			0.062	0.126**	0.122*
			(0.064)	(0.062)	(0.062)
Father Education				0.820***	0.826***
				(0.066)	(0.066)
County Unemployment Rate					0.016*
					(0.009)
Observations	3,796	3,796	3,796	3,796	3,796
R^2	0.007	0.230	0.230	0.260	0.261
Adjusted R^2	0.007	0.229	0.229	0.260	0.260
Residual Std. Error	1.807	1.592	1.592	1.561	1.561
F Statistic	28.476***	565.986***	377.637***	333.487***	267.499***
Note:					*p<0.1; **p<0.05; ***p<0.01

Compared to bivariate regression, all multivariate estimations show a reduction in magnitude of beta_dist. This is evidence that there was OVB and the assumption of exogeneity was not satisfied in the bivariate model.

The R2 greatly improves after adding bytest and dadcoll to the regression

Adding cue80 didn't cause any change in the estimated value of beta_dist which means that we can probably exclude this variable from the regression. Then, we'll continue our analysis assuming that regression (4) correctly controlled for OVB.

Therefore the least biased regression could be said as regression (4).

1.5

Starting with model (4), we have to check if imperfect multicollienarity is an issue:

```r
# Computing VIF for model (4)

# Running auxiliary regressions
aux1_mv4 <- lm(dist ~ bytest + urban + dadcoll, data = colgdata)
aux2_mv4 <- lm(bytest ~ dist + urban + dadcoll, data = colgdata)
aux3_mv4 <- lm(urban ~ dist + bytest + dadcoll, data = colgdata)
aux4_mv4 <- lm(dadcoll ~ dist + bytest + urban, data = colgdata)

# Getting r2
aux1_r2 <- summary(aux1_mv4)$r.squared
aux2_r2 <- summary(aux2_mv4)$r.squared
aux3_r2 <- summary(aux3_mv4)$r.squared
aux4_r2 <- summary(aux4_mv4)$r.squared

# Computing VIF
aux1_vif <- 1 / (1 - aux1_r2)
aux2_vif <- 1 / (1 - aux2_r2)
aux3_vif <- 1 / (1 - aux3_r2)
aux4_vif <- 1 / (1 - aux4_r2)

vifs <- c(aux1_vif, aux2_vif, aux3_vif, aux4_vif)
vifs
```

```
## [1] 1.122645 1.082981 1.121515 1.088459
```

```r
# Testing if VIF are greater than 10
vifs > 10
```

```
## [1] FALSE FALSE FALSE FALSE
```

```r
# Testing if VIF are greater than 5
vifs > 5
```

```
## [1] FALSE FALSE FALSE FALSE
```

Because for all regressors VIF is less than 5 we can be confident that imperfect multicollienarity is not an issue in regression (4). And, if is not an issue in regression (4) - which includes the larger number of independent variables - then it won't be an issue in models (1) to (3)

2.1

$$H_0 : \beta_{\text{Wght}} = 0$$
$$H_a : \beta_{\text{Wght}} \neq 0$$

```
bpdata<- read.csv("bloodpress.csv", stringsAsFactors = FALSE)
Regression1<- lm(BP~Weight, data = bpdata)
stargazer(Regression1,
          title= "Bivariate Regression Results",
          dep.var.labels = c("Blood Pressure"),
          covariate.labels = c("Intercept","Weight (kgs)"),
          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 23, 2019 - 23:05:03

Table 4: Bivariate Regression Results

	Dependent variable:
	Blood Pressure
Intercept	2.205
	(8.663)
Weight (kgs)	1.201***
	(0.093)
Observations	20
R^2	0.903
Adjusted R^2	0.897
Residual Std. Error	1.740 (df = 18)
F Statistic	166.859*** (df = 1; 18)
Note:	*p<0.1; **p<0.05; ***p<0.01

There is a statistically different than 0 relation between Weight and Blood Pressure. If Weight increases by a unit, Blood Pressure increases by 1.201 R^2 value suggests that Change in weight accounts for 89.7% of change in Blood Pressure.

2.2

```
corData1 <- cor(bpdata, use = "pairwise.complete.obs")
corData1 <- corData1[row.names(corData1) %in% c("BP",
                                                "Age",
                                                "Weight",
                                                "BSA",
                                                "Dur",
                                                "Pulse",
                                                "Stress"),
```

```r
                      colnames(corData1) %in% c("BP", "Weight")]
colnames(corData1) <- c("Blood Pressure","Weight")
row.names(corData1) <- c("Blood Pressure",
                        "Age",
                        "Weight",
                        "Body Surface Area",
                        "Duration of Hypertension",
                        "Basal Pulse",
                        "Stress Index")
stargazer(corData1, summary = FALSE,
          type = "latex",
          title = "Correlation Table",
          digits = 2)
```

% 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 23, 2019 - 23:05:03

Table 5: Correlation Table

	Blood Pressure	Weight
Blood Pressure	1	0.95
Age	0.66	0.41
Weight	0.95	1
Body Surface Area	0.87	0.88
Duration of Hypertension	0.29	0.20
Basal Pulse	0.72	0.66
Stress Index	0.16	0.03

```r
Regression2<- lm(BP~Weight + Age, data = bpdata)
Regression3<- lm(BP~Weight + Age + BSA, data = bpdata)
Regression4<- lm(BP~Weight + Age + BSA + Dur, data = bpdata)
Regression5<- lm(BP~Weight + Age + BSA + Dur + Pulse, data = bpdata)
Regression6<- lm(BP~Weight + Age + BSA + Dur + Pulse + Stress, data = bpdata)
Regressions<- list(Regression1,Regression2,Regression3,Regression4,Regression5,Regression6)
stargazer(Regressions,
          title = "Regression Results",
          dep.var.labels = c("Blood Pressure"),
          covariate.labels = c("Intercept",
                               "Weight",
                               "Age",
                               "Body Surface Area",
                               "Duration of Hypertension",
                               "Basal Pulse",
                               "Stress Index"),
          type = "latex",
          intercept.bottom = FALSE,
          df= 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 23, 2019 - 23:05:03

Table 6: Regression Results

	Dependent variable:					
	Blood Pressure					
	(1)	(2)	(3)	(4)	(5)	(6)
Intercept	2.205	-16.579^{***}	-13.667^{***}	-12.852^{***}	-13.127^{***}	-12.870^{***}
	(8.663)	(3.007)	(2.647)	(2.648)	(2.699)	(2.557)
Weight	1.201^{***}	1.033^{***}	0.906^{***}	0.897^{***}	0.926^{***}	0.970^{***}
	(0.093)	(0.031)	(0.049)	(0.048)	(0.060)	(0.063)
Age		0.708^{***}	0.702^{***}	0.683^{***}	0.705^{***}	0.703^{***}
		(0.054)	(0.044)	(0.045)	(0.052)	(0.050)
Body Surface Area			4.627^{***}	4.860^{***}	4.364^{**}	3.776^{**}
			(1.521)	(1.492)	(1.628)	(1.580)
Duration of Hypertension				0.067	0.076	0.068
				(0.049)	(0.051)	(0.048)
Basal Pulse					-0.037	-0.084
					(0.045)	(0.052)
Stress Index						0.006
						(0.003)
Observations	20	20	20	20	20	20
R^2	0.903	0.991	0.995	0.995	0.995	0.996
Adjusted R^2	0.897	0.990	0.994	0.994	0.994	0.994
Residual Std. Error	1.740	0.533	0.437	0.426	0.431	0.407
F Statistic	166.859^{***}	978.248^{***}	971.934^{***}	768.014^{***}	600.750^{***}	560.641^{***}
Note:						*p<0.1; **p<0.05; ***p<0.01

2.3

```
# Computing VIF for model (2)

# Running auxiliary regressions
aux1_mv2 <- lm(Weight ~ Age, data = bpdata)
aux2_mv2 <- lm(Age ~ Weight, data = bpdata)

# Getting r2
aux1_r2 <- summary(aux1_mv2)$r.squared
aux2_r2 <- summary(aux2_mv2)$r.squared

# Computing VIF
aux1_vif <- 1 / (1 - aux1_r2)
aux2_vif <- 1 / (1 - aux2_r2)

vifs <- c(aux1_vif, aux2_vif)
vifs
```

```
## [1] 1.198945 1.198945
```

```
# Testing if VIF are greater than 10
vifs > 10
```

```
## [1] FALSE FALSE
```

```
# Testing if VIF are greater than 5
vifs > 5
```

```
## [1] FALSE FALSE
```

```
# Computing VIF for model (3)

# Running auxiliary regressions
aux1_mv3 <- lm(Weight ~ Age + BSA, data = bpdata)
aux2_mv3 <- lm(Age ~ Weight + BSA, data = bpdata)
aux3_mv3 <- lm(BSA ~ Weight + Age, data = bpdata)

# Getting r2
aux1_r2 <- summary(aux1_mv3)$r.squared
aux2_r2 <- summary(aux2_mv3)$r.squared
aux3_r2 <- summary(aux3_mv3)$r.squared

# Computing VIF
aux1_vif <- 1 / (1 - aux1_r2)
aux2_vif <- 1 / (1 - aux2_r2)
aux3_vif <- 1 / (1 - aux3_r2)

vifs <- c(aux1_vif, aux2_vif, aux3_vif)
vifs
```

```
## [1] 4.403645 1.201901 4.286943
```

```r
# Testing if VIF are greater than 10
vifs > 10
```

```
## [1] FALSE FALSE FALSE
```

```r
# Testing if VIF are greater than 5
vifs > 5
```

```
## [1] FALSE FALSE FALSE
```

```r
# Computing VIF for model (4)

# Running auxiliary regressions
aux1_mv4 <- lm(Weight ~ Age + BSA + Dur, data = bpdata)
aux2_mv4 <- lm(Age ~ Weight + BSA + Dur, data = bpdata)
aux3_mv4 <- lm(BSA ~ Weight + Age + Dur, data = bpdata)
aux4_mv4 <- lm(Dur ~ Weight + Age + BSA, data = bpdata)

# Getting r2
aux1_r2 <- summary(aux1_mv4)$r.squared
aux2_r2 <- summary(aux2_mv4)$r.squared
aux3_r2 <- summary(aux3_mv4)$r.squared
aux4_r2 <- summary(aux4_mv4)$r.squared

# Computing VIF
aux1_vif <- 1 / (1 - aux1_r2)
aux2_vif <- 1 / (1 - aux2_r2)
aux3_vif <- 1 / (1 - aux3_r2)
aux4_vif <- 1 / (1 - aux4_r2)

vifs <- c(aux1_vif, aux2_vif, aux3_vif, aux4_vif)
vifs
```

```
## [1] 4.484932 1.320201 4.344272 1.154968
```

```r
# Testing if VIF are greater than 10
vifs > 10
```

```
## [1] FALSE FALSE FALSE FALSE
```

```r
# Testing if VIF are greater than 5
vifs > 5
```

```
## [1] FALSE FALSE FALSE FALSE
```

```r
# Computing VIF for model (5)

# Running auxiliary regressions
aux1_mv5 <- lm(Weight ~ Age + BSA + Dur + Pulse, data = bpdata)
```

```r
aux2_mv5 <- lm(Age ~ Weight + BSA + Dur + Pulse, data = bpdata)
aux3_mv5 <- lm(BSA ~ Weight + Age + Dur + Pulse, data = bpdata)
aux4_mv5 <- lm(Dur ~ Weight + Age + BSA + Pulse, data = bpdata)
aux5_mv5 <- lm(Pulse ~ Weight + Age + BSA + Dur, data = bpdata)

# Getting r2
aux1_r2 <- summary(aux1_mv5)$r.squared
aux2_r2 <- summary(aux2_mv5)$r.squared
aux3_r2 <- summary(aux3_mv5)$r.squared
aux4_r2 <- summary(aux4_mv5)$r.squared
aux5_r2 <- summary(aux5_mv5)$r.squared

# Computing VIF
aux1_vif <- 1 / (1 - aux1_r2)
aux2_vif <- 1 / (1 - aux2_r2)
aux3_vif <- 1 / (1 - aux3_r2)
aux4_vif <- 1 / (1 - aux4_r2)
aux5_vif <- 1 / (1 - aux5_r2)

vifs <- c(aux1_vif, aux2_vif, aux3_vif, aux4_vif, aux5_vif)
vifs
```

```
## [1] 6.894460 1.762209 5.052259 1.224522 2.986851
```

```r
# Testing if VIF are greater than 10
vifs > 10
```

```
## [1] FALSE FALSE FALSE FALSE FALSE
```

```r
# Testing if VIF are greater than 5
vifs > 5
```

```
## [1]  TRUE FALSE  TRUE FALSE FALSE
```

```r
# Computing VIF for model (6)

# Running auxiliary regressions
aux1_mv6 <- lm(Weight ~ Age + BSA + Dur + Pulse + Stress, data = bpdata)
aux2_mv6 <- lm(Age ~ Weight + BSA + Dur + Pulse + Stress, data = bpdata)
aux3_mv6 <- lm(BSA ~ Weight + Age + Dur + Pulse + Stress, data = bpdata)
aux4_mv6 <- lm(Dur ~ Weight + Age + BSA + Pulse + Stress, data = bpdata)
aux5_mv6 <- lm(Pulse ~ Weight + Age + BSA + Dur + Stress, data = bpdata)
aux6_mv6 <- lm(Stress ~ Weight + Age + BSA + Dur + Pulse, data = bpdata)

# Getting r2
aux1_r2 <- summary(aux1_mv6)$r.squared
aux2_r2 <- summary(aux2_mv6)$r.squared
aux3_r2 <- summary(aux3_mv6)$r.squared
aux4_r2 <- summary(aux4_mv6)$r.squared
aux5_r2 <- summary(aux5_mv6)$r.squared
aux6_r2 <- summary(aux6_mv6)$r.squared
```

```r
# Computing VIF
aux1_vif <- 1 / (1 - aux1_r2)
aux2_vif <- 1 / (1 - aux2_r2)
aux3_vif <- 1 / (1 - aux3_r2)
aux4_vif <- 1 / (1 - aux4_r2)
aux5_vif <- 1 / (1 - aux5_r2)
aux6_vif <- 1 / (1 - aux6_r2)

vifs <- c(aux1_vif, aux2_vif, aux3_vif, aux4_vif, aux5_vif, aux6_vif)
vifs
```

```
## [1] 8.417035 1.762807 5.328751 1.237309 4.413575 1.834845
```

```r
# Testing if VIF are greater than 10
vifs > 10
```

```
## [1] FALSE FALSE FALSE FALSE FALSE FALSE
```

```r
# Testing if VIF are greater than 5
vifs > 5
```

```
## [1]  TRUE FALSE  TRUE FALSE FALSE FALSE
```

2.4

As per the multicollinearity tests and R^2 values, we can infer that regressions (1) to (4) are appropriate to test the original hypothesis. However, regression (4) controls for the OVB from regression (1) and is the least biased of the lot.