INSH5301 Intro Computational Statistics

Ali Banijamali

03/30/2020

College Distance (Re-Re-Revisited):

For this problem we are going to explore the effect of distance from college on educational attainment. The dependent variable (Y) is years of completed education ed. Run all the regressions using the lm command and display regression tables using the stargazer command.

```
# Required Packages:
library('stargazer')

# Reading the data:
col.dist <- read.csv(
  paste('C:/Users/alibs/Google Drive/Courses/INSH5301 Intro Computational Statistics/',
  'Module 10 - Advanced Regression Methods/collegeDistance/CollegeDistance.csv', sep=''),
  stringsAsFactors = F)
```

1. Run a regression of ed on dist, female, bytest, tuition, black, hispanic, incomehi, ownhome, dadcoll, momcoll, cue80, and, stwmfg80. If dist increase from 2 to 3, how are years of education expected to change? If dist increases from 6 to 7, how are years of education expected to change?

ANS.

```
ed.vs.params_1 <- lm(ed ~ dist + female + bytest + tuition + black + hispanic +
                     incomehi + ownhome + dadcoll + momcoll + cue80 + stwmfg80,
                     data=col.dist)

stargazer(ed.vs.params_1, align=TRUE, no.space=TRUE, header=FALSE,
          title="Educational Attainment (Years) vs. Other Parameters (1)",
          dep.var.labels=c("Educational Attainment (Years)"),
          omit.stat=c("LL","ser","f"),
          table.placement = "H" # Hold the table at it's place (Not floating)
         )
```

Table 1: Educational Attainment (Years) vs. Other Parameters (1)

	Dependent variable:
	Educational Attainment (Years)
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)
Constant	8.921^{***}
	(0.252)
Observations	3,796
R^2	0.284
Adjusted R^2	0.281
Note:	*p<0.1; **p<0.05; ***p<0.01

According to the results above, the coefficient for dist is -0.037, therefore a unit increase in distance, has an effect of -0.037 decrease on education years. It is still the case for 6 to 7. These are both one unit increase and a unit increase in distance, changes education years as much as $1 \times (-0.037)$ years.

2. Run a regression of the natural logarithm of ed on dist, female, bytest, tuition, black, hispanic, incomehi, ownhome, dadcoll, momcoll, cue80, and, Stwmfg80. If dist increase from 2 to 3, how are years of education expected to change? If dist increases from 6 to 7, how are years of education expected to change?

ANS.

```
log_ed.vs.params_1 <- lm( I(log(ed)) ~ dist + female + bytest + tuition + black +
                          hispanic + incomehi + ownhome + dadcoll + momcoll +
                          cue80 + stwmfg80,
                          data=col.dist)
```

```
stargazer(log_ed.vs.params_1, align=TRUE, no.space=TRUE, header=FALSE,
          title="LOG(Educational Attainment (Years)) vs. Other Parameters (1)",
          dep.var.labels=c("LOG(Educational Attainment (Years))"),
          omit.stat=c("LL","ser","f"),
          table.placement = "H" # Hold the table at it's place (Not floating)
         )
```

Table 2: LOG(Educational Attainment (Years)) vs. Other Parameters (1)

	Dependent variable:
	LOG(Educational Attainment (Years))
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)
Constant	2.266^{***}
	(0.018)
Observations	3,796
R^2	0.285
Adjusted R^2	0.283
Note:	*p<0.1; **p<0.05; ***p<0.01

The relationship between the dependant and independant variables is log-linear here. According to the results of regression above, the relationship looks like this (I disregarded other independant variables and just mentioned distance):

$\log(ed) = -0.003 \times distance + \beta_i \times x_i$

Let's see how increasing one unit of distance affects ed:

$\log(y_{distance+1}) - \log(y_{distance}) = [-0.003 \times (x+1) + \beta_i \times x_i] - [-0.003 \times (x) + \beta_i \times x_i]$

Therefore:

$\log(\frac{y_{distance+1}}{y_{distance}}) = -0.003 \implies \frac{y_{distance+1}}{y_{distance}} = e^{-0.003}$

This means that for each unit increase in distance (doesn't matter from which distance to which distance, only the difference matters.), ed will be multiplied by $e^{-0.003}$.

3. Run a regression of the natural log ed on dist, dist^2, female, bytest, tuition, black, hispanic, incomehi, ownhome, dadcoll, momcoll, cue80, and, Stwmfg80. If dist increase from 2 to 3, how are years of education expected to change? If dist increases from 6 to 7, how are years of education expected to change?

ANS.

```
log_ed.vs.params_2 <- lm( I(log(ed)) ~ dist + I(dist^2) + female + bytest + tuition +
                          black + hispanic + incomehi + ownhome + dadcoll + momcoll +
                          cue80 + stwmfg80,
                          data=col.dist)

stargazer(log_ed.vs.params_2, align=TRUE, no.space=TRUE, header=FALSE,
          title="LOG(Educational Attainment (Years)) vs. Other Parameters (2)",
          dep.var.labels=c("LOG(Educational Attainment (Years))"),
          omit.stat=c("LL","ser","f"),
          table.placement = "H" # Hold the table at it's place (Not floating)
         )
```

Table 3: LOG(Educational Attainment (Years)) vs. Other Parameters (2)

	Dependent variable:
	LOG(Educational Attainment (Years))
dist	-0.006^{***}
	(0.002)
I(distˆ2)	0.0004^{**}
	(0.0002)
female	0.010^{***}
	(0.004)
bytest	0.007^{***}
	(0.0002)
tuition	-0.014^{**}
	(0.007)
black	0.025^{***}
	(0.005)
hispanic	0.024^{***}
	(0.006)
incomehi	0.026^{***}
	(0.004)
ownhome	0.010^{**}
	(0.005)
dadcoll	0.040^{***}
	(0.005)
momcoll	0.027^{***}
	(0.006)
cue80	0.002^{***}
	(0.001)
stwmfg80	-0.003^{**}
	(0.001)
Constant	2.273^{***}
	(0.018)
Observations	3,796
R^2	0.286
Adjusted R^2	0.284
Note:	*p<0.1; **p<0.05; ***p<0.01

Let's look at the relationship again like the previous question:
$\log(ed) = -0.006d + 0.0004d^2 + \beta_i \times x_i;\ (d : distance)$
$\log(y_{d+1}) - \log(y_d) = [-0.006(d+1) + 0.0004(d+1)^2 + \beta_i \times x_i] - [-0.006d + 0.0004d^2 + \beta_i \times x_i]$
Therefore:
$\log(\frac{y_{d+1}}{y_d}) = -0.0064 + 0.0008d \ => \ \frac{y_{d+1}}{y_d} = e^{-0.0064+0.0008d}$

It can be seen here that there is now a dependency on how distance is changing. To be more specific, it is now important that the unit increase in the distance is from which value to which value. The years of the education will be multiplied by $e^{-0.0064+0.0008d}$ with each unit increase in distance from a distance d. Therefore for 2 to 3 and 6 to 7, these values will be $e^{-0.0064+0.0008(2)} = e^{0.008}$ and $e^{-0.0064+0.0008(6)} = e^{0.0112}$ respectively.

4. Compare the results of (2), and, (3) to (1). Do you think (1) is a correct specification? (Explain)

ANS.
Let's look at the detailed summaries of the 3 model:

```
# I am using summary() instead of stargazer() to see more details of the models.

# Model 1: linear-linear:
summary(ed.vs.params_1)
```

```
##
## Call:
## lm(formula = ed ~ dist + female + bytest + tuition + black +
##     hispanic + incomehi + ownhome + dadcoll + momcoll + cue80 +
##     stwmfg80, data = col.dist)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.3059 -1.1412 -0.2256  1.1672  5.0710
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.920822   0.251586  35.458  < 2e-16 ***
## dist        -0.036661   0.012715  -2.883  0.00396 **
## female       0.142974   0.050437   2.835  0.00461 **
## bytest       0.093038   0.003182  29.242  < 2e-16 ***
## tuition     -0.191052   0.100880  -1.894  0.05832 .
## black        0.350610   0.071230   4.922 8.92e-07 ***
## hispanic     0.361765   0.077270   4.682 2.94e-06 ***
## incomehi     0.371830   0.060721   6.124 1.01e-09 ***
## ownhome      0.138548   0.066723   2.076  0.03792 *
## dadcoll      0.570971   0.073695   7.748 1.19e-14 ***
## momcoll      0.377810   0.081525   4.634 3.70e-06 ***
## cue80        0.028675   0.009866   2.907  0.00368 **
## stwmfg80    -0.042500   0.020208  -2.103  0.03552 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.538 on 3783 degrees of freedom
## Multiple R-squared:  0.2836, Adjusted R-squared:  0.2813
## F-statistic: 124.8 on 12 and 3783 DF,  p-value: < 2.2e-16
```

```
# Model 2: log-linear:
summary(log_ed.vs.params_1)
```

```
##
## Call:
## lm(formula = I(log(ed)) ~ dist + female + bytest + tuition +
##     black + hispanic + incomehi + ownhome + dadcoll + momcoll +
##     cue80 + stwmfg80, data = col.dist)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.30987 -0.08149 -0.01438  0.08585  0.34586
##
```

```
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.2658193  0.0178617 126.854  < 2e-16 ***
## dist        -0.0026072  0.0009027  -2.888  0.00390 **
## female       0.0103059  0.0035808   2.878  0.00402 **
## bytest       0.0066561  0.0002259  29.467  < 2e-16 ***
## tuition     -0.0139382  0.0071621  -1.946  0.05172 .
## black        0.0261676  0.0050571   5.174 2.40e-07 ***
## hispanic     0.0259986  0.0054859   4.739 2.22e-06 ***
## incomehi     0.0265197  0.0043110   6.152 8.47e-10 ***
## ownhome      0.0098332  0.0047371   2.076  0.03798 *
## dadcoll      0.0405374  0.0052321   7.748 1.19e-14 ***
## momcoll      0.0266016  0.0057880   4.596 4.45e-06 ***
## cue80        0.0020357  0.0007004   2.906  0.00368 **
## stwmfg80    -0.0028642  0.0014347  -1.996  0.04597 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1092 on 3783 degrees of freedom
## Multiple R-squared:  0.2853, Adjusted R-squared:  0.283
## F-statistic: 125.9 on 12 and 3783 DF,  p-value: < 2.2e-16
```

```
# Model 3: log-quadratic:
summary(log_ed.vs.params_2)
```

```
##
## Call:
## lm(formula = I(log(ed)) ~ dist + I(dist^2) + female + bytest +
##     tuition + black + hispanic + incomehi + ownhome + dadcoll +
##     momcoll + cue80 + stwmfg80, data = col.dist)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.31213 -0.08088 -0.01339  0.08528  0.34592
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.2728899  0.0181423 125.281  < 2e-16 ***
## dist        -0.0060526  0.0018137  -3.337 0.000854 ***
## I(dist^2)    0.0003593  0.0001640   2.190 0.028587 *
## female       0.0103322  0.0035790   2.887 0.003913 **
## bytest       0.0066250  0.0002262  29.286  < 2e-16 ***
## tuition     -0.0140750  0.0071588  -1.966 0.049357 *
## black        0.0248766  0.0050888   4.888 1.06e-06 ***
## hispanic     0.0237960  0.0055746   4.269 2.01e-05 ***
## incomehi     0.0263391  0.0043096   6.112 1.09e-09 ***
## ownhome      0.0101988  0.0047377   2.153 0.031405 *
## dadcoll      0.0397778  0.0052409   7.590 4.01e-14 ***
## momcoll      0.0265932  0.0057851   4.597 4.43e-06 ***
## cue80        0.0018250  0.0007067   2.583 0.009844 **
## stwmfg80    -0.0028683  0.0014340  -2.000 0.045545 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1091 on 3782 degrees of freedom
```

```
## Multiple R-squared:  0.2862, Adjusted R-squared:  0.2838
## F-statistic: 116.7 on 13 and 3782 DF,  p-value: < 2.2e-16
```

I don't see a major benefit in models 2 and 3 over 1. All 3 models have a p-value of $<$ 2.2e-16. Model 2 is just a little better than 1: The adjusted R^2 barely changed in model 2 (from 0.2813 in model 1 to 0.2830 in model 2), and model 3 is just slightly better than model 2 with and adjusted R^2 of 0.2838.

5. Add the interaction term dadcoll*momcoll to the regression (3). Interpret the coefficient of the interaction term.

ANS.

```
log_ed.vs.params_3 <- lm(I( log(ed) ) ~ dist + I(dist^2) + female + bytest + tuition +
                         black + hispanic + incomehi + ownhome + dadcoll + momcoll +
                         cue80 + stwmfg80 + dadcoll*momcoll,
                         data=col.dist)

stargazer(log_ed.vs.params_3, align=TRUE, no.space=TRUE, header=FALSE,
          title="LOG(Educational Attainment (Years)) vs. Other Parameters (3)",
          dep.var.labels=c("LOG(Educational Attainment (Years))"),
          omit.stat=c("LL","ser","f"),
          table.placement = "H" # Hold the table at it's place (Not floating)
         )
```

Table 4: LOG(Educational Attainment (Years)) vs. Other Parameters (3)

	Dependent variable:
	LOG(Educational Attainment (Years))
dist	-0.006^{***}
	(0.002)
I(distˆ2)	0.0004^{**}
	(0.0002)
female	0.010^{***}
	(0.004)
bytest	0.007^{***}
	(0.0002)
tuition	-0.014^{**}
	(0.007)
black	0.025^{***}
	(0.005)
hispanic	0.024^{***}
	(0.006)
incomehi	0.026^{***}
	(0.004)
ownhome	0.010^{**}
	(0.005)
dadcoll	0.047^{***}
	(0.006)
momcoll	0.041^{***}
	(0.008)
cue80	0.002^{**}
	(0.001)
stwmfg80	-0.003^{*}
	(0.001)
dadcoll:momcoll	-0.027^{**}
	(0.011)
Constant	2.272^{***}
	(0.018)
Observations	3,796
R^2	0.287
Adjusted R^2	0.285
Note:	*p<0.1; **p<0.05; ***p<0.01

The interaction effect is interesting! While father and mother college education, each one independently have positive effects (coeff's are +0.047 and +0.041 respectively), meaning that if each one of the parents have college education, that leads to higher years of education, the interaction effect is negative! (coeff = -0.027), meaning that if both of the parents have college education the total length of education decreases.

6. Test if the effect of distance on education depends on the family's income.

ANS.
For this question we can compare to exactly similar models containng distance, with and without family income variable:

```r
# 1. Model w/ only distance:
ed.vs.dist <- lm(ed ~ dist, data=col.dist)
summary(ed.vs.dist)
```

```
##
## Call:
## lm(formula = ed ~ dist, data = col.dist)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9559 -1.8091 -0.6624  2.0515  4.4844
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.95586    0.03772 369.945   <2e-16 ***
## dist        -0.07337    0.01375  -5.336    1e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.807 on 3794 degrees of freedom
## Multiple R-squared:  0.00745,    Adjusted R-squared:  0.007188
## F-statistic: 28.48 on 1 and 3794 DF,  p-value: 1.004e-07
```

```r
# 2. Model w/ distance + family income:
ed.vs.dist_income <- lm(ed ~ dist + incomehi, data=col.dist)
summary(ed.vs.dist_income)
```

```
##
## Call:
## lm(formula = ed ~ dist + incomehi, data = col.dist)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.5337 -1.6000 -0.5221  1.4954  4.6619
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.68735    0.04205 325.508  < 2e-16 ***
## dist        -0.05821    0.01349  -4.315 1.64e-05 ***
## incomehi     0.84635    0.06367  13.293  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.767 on 3793 degrees of freedom
## Multiple R-squared:  0.05163,    Adjusted R-squared:  0.05113
## F-statistic: 103.2 on 2 and 3793 DF,  p-value: < 2.2e-16
```

The effect of distance barely changed after introducing family income to the model (β changed from -0.07337 to -0.05821) and in both models it was significant so I don't think that the effect of distance depends very much on family income.

7. Use the F-test on (5) as the complete model and (2) as the nested model. What specification do you prefer? (Explain)

ANS.
From the slides of module 9.3:
To test if a set of variables significantly contribute to the regression the solution is to setup an F test. Here, we are not testing all the variables at once, instead, we are testing one set of variables (the complete model) against a subset of variables (the reduced model, where we have droped 1 or more variables).
The null hypothesis is that the complete model does no better than the reduced model (ie, that the variables you are debating including are not significant; this is the same null as we have for a single variable when we examine the t statistic on its beta coefficient). And as usual, if we get a F statistic that is sufficiently large, then we conclude that the complete model with the extra variables is signficantly better than the reduced model in explaining y, and thus we are justified in including all the variables under debate.

We can setup this test in R by doing an F test on the complete model vs the reduced model:

```
# Here, our complete model is: log_ed.vs.params_3
# and our reduced model is: log_ed.vs.params_1
anova(log_ed.vs.params_1, log_ed.vs.params_3)
```

```
## Analysis of Variance Table
##
## Model 1: I(log(ed)) ~ dist + female + bytest + tuition + black + hispanic +
##     incomehi + ownhome + dadcoll + momcoll + cue80 + stwmfg80
## Model 2: I(log(ed)) ~ dist + I(dist^2) + female + bytest + tuition + black +
##     hispanic + incomehi + ownhome + dadcoll + momcoll + cue80 +
##     stwmfg80 + dadcoll * momcoll
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1   3783 45.091
## 2   3781 44.967  2   0.12434 5.2277 0.005405 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

According to the significance level (0.005405) we can say that the complete model with the additional parameters is better.