Problem Set 2 - Solutions

Statistics and Econometrics

Question 1

Use bwght.RData for this exercise. The data set contains data on births to women in the United States. A problem of interest to health officials (and others) is to determine the effects of smoking during pregnancy on infant health. One measure of infant health is birth weight (variable *bwght*); a birth weight that is too low can put an infant at risk for contracting various illnesses. Since factors other than cigarette smoking (variable *cigs*) that affect birth weight are likely to be correlated with smoking, we should take those factors into account. For example, higher income (variable $faminc$) generally results in access to better prenatal care, as well as better nutrition for the mother. An equation that recognizes this is:

$$bwght = \beta_0 + \beta_1 cigs + \beta_2 faminc + u.$$

1. What is the most likely sign for β_2?
2. Do you think *cigs* and $faminc$ are likely to be correlated? Explain why the correlation might be positive or negative.
3. Now, estimate the equation with and without $faminc$. Report the results. Discuss your results, focusing on whether adding $faminc$ substantially changes the estimated effect of *cigs* on *bwght*.

Solutions

1. Probably $\beta_2 > 0$, as more income typically means better nutrition for the mother and better prenatal care.
2. On the one hand, an increase in income makes cigarettes more affordable, and *cigs* and $faminc$ could be positively correlated. On the other, family incomes are also higher for families with more education, and more education and cigarette smoking tend to be negatively correlated.

```
load("bwght.RData")
cor(data$cigs, data$faminc)
```

```
## [1] -0.1730449
```

The sample correlation between *cigs* and $faminc$ is about -.173, indicating a negative correlation.

3.

```
bwght.m1 <- lm(bwght ~ cigs, data = data)
bwght.m2 <- lm(bwght ~ cigs + faminc, data = data)
stargazer(bwght.m1, bwght.m2, header = FALSE, type = 'latex', title = "Question 1.3")
```

The effect of cigarette smoking is slightly smaller (i.e., the coefficient of *cigs* increases) when $faminc$ is added to the regression, but the difference is not great. This is due to the fact that *cigs* and $faminc$ are not very correlated, and the coefficient on $faminc$ is practically small. (The variable $faminc$ is measured in thousands, so $10,000 more in 1988 income increases predicted birth weight by only .93 ounces.)

Question 2

The following model can be used to study whether campaign expenditures affect election outcomes:

$$voteA = \beta_0 + \beta_1 \log(expendA) + \beta_2 \log(expendB) + \beta_3 prtystrA + u,$$

Table 1: Question 1.3

	Dependent variable:	
	bwght	
	(1)	(2)
cigs	-0.514^{***}	-0.463^{***}
	(0.090)	(0.092)
faminc		0.093^{***}
		(0.029)
Constant	119.772^{***}	116.974^{***}
	(0.572)	(1.049)
Observations	1,388	1,388
R^2	0.023	0.030
Adjusted R^2	0.022	0.028
Residual Std. Error	20.129 (df = 1386)	20.063 (df = 1385)
F Statistic	32.235^{***} (df = 1; 1386)	21.274^{***} (df = 2; 1385)
Note:		*p<0.1; **p<0.05; ***p<0.01

where $voteA$ is the percentage of the vote received by Candidate A, $expendA$ and $expendB$ are campaign expenditures by Candidate A and B, and $prtystrA$ is a measure of party strength for Candidate A (the percentage of the most recent presidential vote that went to A's party).

1. What is the interpretation of β_1?
2. Estimate the given model using vote1.RData and report the results. Do A's expenditures affect the outcome? What about B's expenditures? (please show clearly the test statistic and the critical value used in your testing).
3. In terms of the parameters, state the null hypothesis that a 1% increase in A's expenditures is offset by a 1% increase in B's expenditures.
4. Test the hypothesis in part 3. What is your conclusion?

Solutions

1. Holding other factors fixed,

$$\Delta voteA = \beta_1 \Delta \log(expendA) \approx (\beta_1/100)(\%\Delta expendA).$$

   So $\beta_1/100$ is the pecentage point change in $voteA$ when $expendA$ increases by one percent.

2.

```
load("vote1.RData")
vote.m1 <- lm(voteA ~ log(expendA) + log(expendB) + prtystrA, data = data)
stargazer(vote.m1, header = FALSE, type = 'latex', title = "Question 2.2")
```

The t statistic of coefficient on log($expendA$) is approximately 15.92, and the t statistic of coefficient on log($expendB$) is approximately -17.45. The critical value associated with the 99.5% percentile in a standard normal distribution is around 2.576. So both coefficients are statistically significant at 1% level. The estimates imply that a 10% increase in spending by candidate A increases the predicted share of the vote going to

Table 2: Question 2.2

	Dependent variable:
	voteA
log(expendA)	6.083***
	(0.382)
log(expendB)	−6.615***
	(0.379)
prtystrA	0.152**
	(0.062)
Constant	45.079***
	(3.926)
Observations	173
R^2	0.793
Adjusted R^2	0.789
Residual Std. Error	7.712 (df = 169)
F Statistic	215.227*** (df = 3; 169)
Note:	*p<0.1; **p<0.05; ***p<0.01

A by about .61 percentage points. Similarly, a 10% increase in spending by B reduces voteA by about .66 percentage points. These effects certainly cannot be ignored.

3. The null hypothesis is $H_0 : \beta_2 = -\beta_1$, which means a z% increase in expenditure by A and a z% increase in expenditure by B leaves voteA unchanged. We can equivalently write $H_0 : \beta_1 + \beta_2 = 0$.

4.

```r
linearHypothesis(vote.m1, "log(expendA) + log(expendB) = 0")
```

```
## Linear hypothesis test
##
## Hypothesis:
## log(expendA)  + log(expendB) = 0
##
## Model 1: restricted model
## Model 2: voteA ~ log(expendA) + log(expendB) + prtystrA
##
##   Res.Df   RSS Df Sum of Sq      F Pr(>F)
## 1    170 10111
## 2    169 10052  1    59.261 0.9963 0.3196
```

The p value for the test is as high as 0.3196, so we fail to reject $H_0 : \beta_2 = -\beta_1$.

Question 3

In class, we discussed the concepts of individual significance (Slide 15) and joint significance (Slide 36). Is it possible that a variable is individually significant but when we test the joint significance of this variable, along with some other variables, we find them jointly insignificant? Explain.

Solutions

Yes, it is possible. Consider a model like this:

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

Suppose x_1 is significant at certain significance level, while others are irrelevant variables, meaning that $\beta_2 = \beta_3 = \cdots = \beta_k = 0$.

Significance of x_1 means that, if we run an F test for the null hypothesis $H_0 : \beta_1 = 0$, the corresponding F statistic $F_1 = \frac{(SSR_r - SSR_{ur})/1}{SSR_{ur}/(n-k-1)}$ is larger than the critical value.

Now think about the case when we want to test the joint significance of x_1 and x_2, i.e., the null hypothesis is $H_0 : \beta_1 = 0, \beta_2 = 0$. The corresponding F statistic is $F_2 = \frac{(SSR'_r - SSR_{ur})/2}{SSR_{ur}/(n-k-1)}$. It is easy to see that $SSR'_r = SSR_r$ because x_2 is irrelavant, and thus F_2 is equal to half of F_1. Following this logic, when we want to test the joint significance of x_1, along with other $k-1$ irrelevant variables, F_1 shrinks by a factor of k. When k is sufficiently large, the F statistic is guaranteed to be smaller than any critical value associated with the common choice of significance levels, implying joint insignificance.