## Exercise Session 5

### Problem 1

Suppose that you have a sample of $n$ individuals who apart from their mother tongue (Czech) can speak English, German, or are trilingual (i.e., all individuals in your sample speak in addition to their mother tongue at least one foreign language). You estimate the following model:

$$wage = \beta_0 + \beta_1 educ + \beta_2 IQ + \beta_3 exper + \beta_4 DM + \beta_5 Germ + \beta_6 Engl + \varepsilon \ ,$$

where

| | | |
|---|---|---|
| $educ$ | . . . | years of education |
| $IQ$ | . . . | IQ level |
| $exper$ | . . . | years of on-the-job experience |
| $DM$ | . . . | dummy, equal to one for males and zero for females |
| $Germ$ | . . . | dummy, equal to one for German speakers and zero otherwise |
| $Engl$ | . . . | dummy, equal to one for English speakers and zero otherwise |

a. Explain why a dummy equal to one for trilingual people and zero otherwise is not included in the model.

**If we included the dummy for people who are trilingual, we would have the complete set of dummies in the model (describing all three possible options - German speaker, English speaker, both foreign languages). Since we have the intercept in the model, this would lead to perfect multicollinearity.**

b. Explain how you would test for discrimination against females (in the sense that *ceteris paribus* females earn less than males). Be specific: state the hypothesis, give the test statistic and its distribution.

**For women, the dummy $DM$ is equal to 0 and the model stands as follows:**

$$\boldsymbol{wage = \beta_0 + \beta_1 educ + \beta_2 IQ + \beta_3 exper + \beta_5 Germ + \beta_6 Engl + \varepsilon \ .}$$

**For men, the dummy $DM$ is equal to 1 and the model stands as follows:**

$$\boldsymbol{wage = \beta_0 + \beta_1 educ + \beta_2 IQ + \beta_3 exper + \beta_4 + \beta_5 Germ + \beta_6 Engl + \varepsilon \ .}$$

**Therefore, ceteris paribus, the difference between the wage of men and the wage of women is equal to $\beta_4$. If this coefficient is positive, then men earn more than women. Hence, our hypothesis to be tested is**

$$H_0 : \beta_4 \leq 0 \ \text{ vs } \ H_A : \beta_4 > 0 \ .$$

**This leads to a one-sided $t$-test with the test statistic**

$$t = \frac{\widehat{\beta_4}}{SE(\widehat{\beta_4})} \sim t_{n-k}$$

**where $k$ = 7 in this case. When we compute this test statistic, we compare it to the critical value $t_{n-7,0.95}$. If the test statistic is larger than this critical value, then we reject the $H_0$ at 95% confidence level and we conclude that there is discrimination against females. where $k$ = 7 in this case. When we compute this test statistic, we compare it to the critical value $t_{n-7,0.95}$. If the test statistic is larger than this critical value, then we reject the $H_0$ at 95% confidence level and we conclude that there is discrimination against females.**

c. Explain how you would measure the payoff (in terms of wage) to someone of becoming trilingual given that he can already speak (i) English, (ii) German.

**The payoff of a trilingual person is**

$$wage = \beta_0 + \beta_1 educ + \beta_2 IQ + \beta_3 exper + \beta_4 DM + \beta_5 + \beta_6 + \varepsilon\ ,$$

**the payoff of a German speaking person is**

$$wage = \beta_0 + \beta_1 educ + \beta_2 IQ + \beta_3 exper + \beta_4 DM + \beta_5 + \varepsilon\ ,$$

**and the payoff of an English speaking person is**

$$wage = \beta_0 + \beta_1 educ + \beta_2 IQ + \beta_3 exper + \beta_4 DM + \beta_6 + \varepsilon\ .$$

**Hence, by becoming trilingual, a person who can already speak English gains $\beta_5$ and a person who can already speak German gains $\beta_6$. If we assume that both coefficients are positive, this payoff should be positive.**

d. Explain how you would test if the influence of on-the-job experience is greater for males than for females. Be specific: specify the model, state the hypothesis, give the test statistic and its distribution.

**To allow the on-the-job experience to be greater for males than for females, we have to define a slope coefficient on *exper* that would be different for males and for females. We can do so using the following model:**

$$wage = \beta_0 + \beta_1 educ + \beta_2 IQ + \beta_3 exper + \beta_4 DM + \beta_5 Germ + \beta_6 Engl + \beta_7 exper \cdot DM + \varepsilon\ .$$

**Where we have created an interaction term exper*DM. In this case, the impact of on the on-the-job experience on wage would be $\beta_3$ for females and $\beta_3 + \beta_7$ for males. Hence, if $\beta_7$ is positive, then men gain more from experience than women. Hence, our hypothesis to be tested is**

$$H_0 : \beta_7 \leq 0 \text{ vs } H_A : \beta_7 > 0\ .$$

$$t = \frac{\widehat{\beta_7}}{SE(\widehat{\beta_7})} \sim t_{n-k}$$

**where $k =$ 8 in this case. When we compute this test statistic, we compare it to the critical value $t_{n-8,0.95}$. If the test statistic is larger than this critical value, then we reject the $H_0$ at 95% confidence level and we conclude that the influence of on-the-job experience is greater for males than for females.**

**Problem 2**

We have information about mortality rates (MORT=total mortality rate per 100,000 population) in a specific year for 51 States of the United States combined with information about potential determinants: INCC (per capita income by State in Dollars), POV (proportion of families living below the poverty line), EDU (proportion of population completing 4 years of high school), TOBC (per capita consumption of cigarettes by State) and AGED (proportion of population over the age of 65). Estimation results are presented in the following table:

**OLS Estimation Results**

| Variable | Model 1 coefficients | Model 2 coefficients | Model 3 coefficients |
|---|---|---|---|
| Constant | 194.747 | 531.608 | -9.231 |
| | (53.915) | (94.409) | (176.795) |
| Aged | 5,546.56 | 5,024.38 | 5,311.4 |
| | (445.727) | (358.218) | (334.415) |
| Incc | | 0.014 | 0.015 |
| | | (0.0038) | (0.0037) |
| Edu | | -682.591 | -285.715 |
| | | (114.812) | (152.926) |
| Pov | | | 854.178 |
| | | | (302.345) |
| Tobc | | | 0.989 |
| | | | (0.342) |
| n | 51 | 51 | 51 |
| Adjusted R squared | 0.759 | 0.856 | 0.884 |
| SSR | 228,770.3 | 128,260.1 | 99,303.73 |

i) Interpret the slope coefficient in Model 1 and validate it at 1% significance level.

**The slope coefficient in Model 1 implies that increasing the proportion of population over the age 65 by 1 percentage points will increase the mortality rate per 100000 of population by 5 5.46 T statistics of the coefficient is t=$\frac{5556.56}{445.727} = 12.46$, which is > 2.7, therefore, it is significant at the 1% significance level**

ii) Validate the joint significance of Model 2 in comparison to model 1 at 1% significance level?

**Joint significance of model 2 in comparison to model 1 implies that we test joint significance of the coefficients Incc and Edu, according to F test we have** $F = \frac{(228770.3-128260.1)/2}{128260.1/47} = \frac{50255.1}{2728.94} = \mathbf{18.41}$ **comparing this with the critical value 5.1 we reject the hypothesis that the coefficients Incc and Edu are jointly insignificant**

iii) Comment on the effect of INCC on MORT in the second model. Why do you think is a positive and significant effect? **When per capita income by state is higher, the mortality rate is also higher. This at one glance does not make any sense because rich people should be able to afford better health care and therefore, extend longevity of their lives. However, we can make an argument that generally, older people are more likely to have higher income, therefore, those states with high income probably also have proportion of old people higher and hence, the mortality rate is higher.**

iv) In Model 3 we add two new explanatory variables: POV and TOBC. Test whether this inclusion helps to improve the quality of the model at 1% significance level. Is model 3 the best in terms of goodness-of-fit? **This question indirectly asks to compare the model 3 to model 2, therefore, we need to test joint significance of the variables POV and TOBC**

**Again we calculate an F test F=** $\frac{(128260.1-99303.73)/2}{99303.73/45} = \frac{14478}{2206} = 6.56$, **which is still larger than the critical value 5.1, therefore, these two variables are jointly significant at the 1% significance level. The model is the best in terms of the goodness of fit, because** $\mathbf{R^{adj}}$ **is the highest**

v) Are the effects of these two new variables the expected ones? Are they individually significant at 1% significance level? **These two new variables have "positive" impact on the mortality rate, which makes sense, more smokers- higher mortality, more poor people – higher mortality. T test for POV is 2.8 and for TOBC, 2.9, both are statistically significant at 1% significance level when comparing to critical value 2.7**

vi) What about the individual significance of EDU in model 3 if compared with model 2? Why? **Edu in model 2 is significant while it is not in model 3. The fact that POV and TOBC were omitted in the model 2 was causing a bias in the estimation of the coefficient Edu. The reason is that EDU is negatively correlated with both POV and TOBC – more educated people are less likely to smoke and less likely to be poor. Meanwhile, POV and TOBC are positively correlated with the explained variable MORT, meaning that the direction of bias is negative. Since coefficient on Edu is negative, this bias was making it more negative in absolute terms and this way it was making it significant.**

## Problem 3

consider a simple model to compare the returns to education at junior colleges and four-year colleges; for simplicity, we refer to the latter as "universities." The population includes working people with a high school degree, and the model is:

$$\log(wage) = \alpha_0 + \alpha_1 jc + \alpha_2 univ + \alpha_3 exper + u \qquad (1)$$

where

*jc* is number of years attending a two-year college, *univ* is number of years at a four-year college. *exper* is months in the workforce.

Note that any combination of junior college and four-year college is allowed, including *jc* =0 and *univ* = 0. Use the data ***twoyear.dta***

a) Test the hypothesis that $\alpha_1 = \alpha_2$. The hypothesis of interest is whether one year at a junior college is worth one year at a university.
**To test this hypothesis we instead want to test $\theta = \alpha_1 - \alpha_2 = 0$ and plug it in the original regression:**

$$\log(wage) = \alpha_0 + (\theta + \alpha_2)jc + \alpha_2 univ + \alpha_3 exper + u$$
$$\log(wage) = \alpha_0 + \theta jc + \alpha_2(univ + jc) + \alpha_3 exper + u \qquad (2)$$

**Now run:**
**genr unjc=univ+jc**
**ols lwage const jc unjc exper**

| | coefficient | std. error | t-ratio | p-value | |
|---|---|---|---|---|---|
| const | 1.47233 | 0.0210602 | 69.91 | 0.0000 | *** |
| jc | -0.0101795 | 0.00693591 | -1.468 | 0.1422 | |
| unjc | 0.0768762 | 0.00230873 | 33.30 | 2.96e-225 | *** |
| exper | 0.00494422 | 0.000157474 | 31.40 | 4.12e-202 | *** |

| | | | |
|---|---|---|---|
| Mean dependent var | 2.248096 | S.D. dependent var | 0.487692 |
| Sum squared resid | 1250.544 | S.E. of regression | 0.430138 |
| R-squared | 0.222442 | Adjusted R-squared | 0.222097 |
| F(3, 6759) | 644.5330 | P-value(F) | 0.000000 |
| Log-likelihood | -3888.687 | Akaike criterion | 7785.374 |
| Schwarz criterion | 7812.651 | Hannan-Quinn | 7794.789 |

$\widehat{\alpha_1} - \widehat{\alpha_2} = -0.0102$ **so the return to a year at a junior college is about one percentage point less than a year at a university.**
**Test statistic on jc t=0.0102/.0069 =-1.48. We need to compare this with one sided alternative critical value. At 10% one-sided significance level, critical value is -1.282. Therefore, there is some but not strong evidence against the null hypothesis.**
**Check also command: ols lwage const jc univ exper. Make your own observation!**

(ii) The variable phsrank is the person's high school percentile. (A higher number is better. For example, 90 means you are ranked better than 90 percent of your graduating class.) Find the smallest, largest, and average phsrank in the sample.
**summary phsrank**
(ii) Add phsrank to regression (2) and report the OLS estimates in the usual form. Is phsrank statistically significant? How much is 10 percentage points of high school rank worth in terms of wage?
**ols lwage const jc unjc exper phsrank**
***phsrank* has a *t* statistic equal to only 1.25; it is not statistically significant. If we increase *phsrank* by 10, log(*wage*) is predicted to increase by (.0003)10 = .003. This implies a .3% increase in *wage*, which seems a modest increase given a 10 percentage point increase in *phsrank*.**
(iii) Does adding phsrank to regression (2) substantively change the conclusions on the returns to two- and four-year colleges? Explain.
**Adding *phsrank* makes the *t* statistic on *jc* even smaller in absolute value, about 1.33, but**

**the coefficient magnitude is similar to (2). Therefore, the base point remains unchanged: the return to a junior college is estimated to be somewhat smaller, but the difference is barely significant with one-sided test.**