

## BUEC 333 QUIZ

Consider the following regression output from EViews, noting that wages give total annual earnings from wages and salaries, agegrp is 5-year age groups, where agegrp=9 is 25-29 years old, hdgree indicates highest degree of schooling in 13 levels, sex=1 for women, citizen=1 for persons who are Canadian-by-birth, vismin indicates visible minority group membership, where values 1-12 are visible minority groups, and aboid indicates Aboriginal identity (including registry) for values 1-5.

Note that the regression output was generated by

QUICK-->EQUATION:

equation specification, using @expand:

```
log(wages) c @expand(agegrp, @dropfirst) @expand(hdgree, @dropfirst)
(vismin<13) (aboid<6) sex=1
```

sample:

```
1 56529 if agegrp>8 and agegrp<17 and wages>100 and citizen=1
```

Dependent Variable: LOG(WAGES)

Method: Least Squares

Date: 11/06/10 Time: 19:25

Sample: 1 56529 IF AGEGRP>8 AND AGEGRP<17 AND WAGES>100

AND CITIZEN=1

Included observations: 13236

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	9.833821	0.038375	256.2582	0.0000
VISMIN<13	0.020050	0.028715	0.698230	0.4850
ABOID<6	-0.212863	0.048369	-4.400816	0.0000
SEX=1	-0.422557	0.016885	-25.02616	0.0000
AGEGRP=10	0.307883	0.030343	10.14672	0.0000
AGEGRP=11	0.417287	0.030683	13.60013	0.0000
AGEGRP=12	0.580113	0.030735	18.87470	0.0000
AGEGRP=13	0.636205	0.030626	20.77356	0.0000
AGEGRP=14	0.618184	0.032097	19.25971	0.0000
AGEGRP=15	0.487574	0.034508	14.12919	0.0000
AGEGRP=16	0.130724	0.042995	3.040436	0.0024
HDGREE=2	0.283840	0.035274	8.046731	0.0000
HDGREE=3	0.239772	0.046314	5.177108	0.0000
HDGREE=4	0.415034	0.050108	8.282826	0.0000
HDGREE=5	0.262411	0.048125	5.452640	0.0000
HDGREE=6	0.388865	0.039400	9.869772	0.0000
HDGREE=7	0.495839	0.047258	10.49211	0.0000
HDGREE=8	0.439606	0.047437	9.267055	0.0000
HDGREE=9	0.622720	0.037117	16.77728	0.0000
HDGREE=10	0.586505	0.058468	10.03122	0.0000
HDGREE=11	1.021631	0.112163	9.108484	0.0000
HDGREE=12	0.760526	0.047171	16.12283	0.0000
HDGREE=13	1.064670	0.100071	10.63910	0.0000
HDGREE=88	0.261332	0.247468	1.056025	0.2910
R-squared	0.126205	Mean dependent var	10.43770	
Adjusted R-squared	0.124684	S.D. dependent var	1.015558	
S.E. of regression	0.950139	Akaike info criterion	2.737395	
Sum squared resid	11927.32	Schwarz criterion	2.750977	
Log likelihood	-18092.08	Hannan-Quinn criter.	2.741929	
F-statistic	82.96733	Durbin-Watson stat	2.124936	
Prob(F-statistic)	0.000000			

- 1) [4 points] Hypothesis Tests (provide reasoning or mathematical steps for all questions). You may assume that the  $t$  distribution with the large a sample size is equal to the standard normal distribution.
- AGEGRP=12 indicates a person aged 40-44. AGEGRP=13 indicates a person aged 45-49. Given these estimates, how much more does woman aged 48 earn than a woman aged 42? What about the same comparison for a man?
  - Test the hypothesis that visible minority workers earn the same as white workers with the same age and education. Test the hypothesis that Aboriginal workers earn the same as white workers.
  - Test the hypothesis that white workers have a log-earnings premium in comparison with Aboriginal workers that is larger than 10 per cent.
  - The  $t$ -statistic for the constant term (C) is very large. Why is it so large?
- 2) [4 points] Other Topics (provide reasoning or mathematical steps for all questions)
- R-squared is 0.126205. Does that mean that 87.3 per cent of the variation of log-earnings is unexplained? Does that mean that the regression is not informative? Does that mean that we should disregard that results of the tests above?
  - Does the inclusion of both ( $vismin<13$ ) and ( $sex=1$ ) induce multicollinearity? If so, how?
  - Suppose that earnings were correlated with field-of-study, with mathematical fields of study having higher earnings than nonmathematical fields of study. Suppose also that visible minorities are more likely to choose mathematical fields of study. Describe formally how this might lead to endogeneity. Describe how it would change your interpretation of the coefficient on ( $vismin<13$ ).
  - Suppose that the data included a variable called  $age$  equal to the exact age in years for each respondent. In that case, would it have been a good idea to drop the age group variables and replace them simply with the one age variable? Why or why not?
- 3) [2 points] Suppose that  $Y_i = \beta_0 + \beta_1 X_i + u_i + v_i$  and that  $u_i$  and  $v_i$  are 2 independently distributed random variables, each of which takes on the value -1 with 50% probability and the value 1 with 50% probability. Using the 6 classical assumptions, suppose that we regress  $Y$  on  $X$ . Using the 6 classical assumptions, show whether or not the estimated OLS coefficients are unbiased. Using the 6 classical assumptions, show whether or not the estimated OLS coefficients are efficient.

### Some probabilities for the standard normal distribution, $z$ .

2 tailed probabilities:  $\text{Prob}[-1.65 < z < 1.65] = 0.90$ ,  $\text{Prob}[-1.96 < z < 1.96] = 0.95$ ,  $\text{Prob}[-2.56 < z < 2.56] = 0.99$

1 tailed probabilities:  $\text{Prob}[z < 1.28] = 0.90$ ,  $\text{Prob}[z < 1.65] = 0.95$ ,  $\text{Prob}[z < 2.33] = 0.99$

### The 6 Classical Assumptions

- The regression model is **linear in the coefficients, correctly specified**, and has an **additive error term**.
- The error term has **zero population mean**:  $E(\varepsilon_i) = 0$ .
- All independent variables are **uncorrelated with the error term**:  $\text{Cov}(X_i, \varepsilon_i) = 0$  for each independent variable  $X_i$
- Errors are uncorrelated across observations:  $\text{Cov}(\varepsilon_i, \varepsilon_j) = 0$  for two observations  $i$  and  $j$  (no **serial correlation**).
- The error term has constant variance:  $\text{Var}(\varepsilon_i) = \sigma^2$  for every  $i$  (no **heteroskedasticity**).
- No independent variable is a **perfect linear function** of any other independent variable (no **perfect multicollinearity**).

### Useful Formulae:

$$\mu_x = E(X) = \sum_{i=1}^k p_i x_i \quad \text{Var}(X) = E\left[(X - \mu_x)^2\right] = \sum_{i=1}^k (x_i - \mu_x)^2 p_i$$

$$\text{Cov}(X, Y) = \sum_{i=1}^k \sum_{j=1}^m (x_j - \mu_x)(y_i - \mu_y) \Pr(X = x_j, Y = y_i)$$