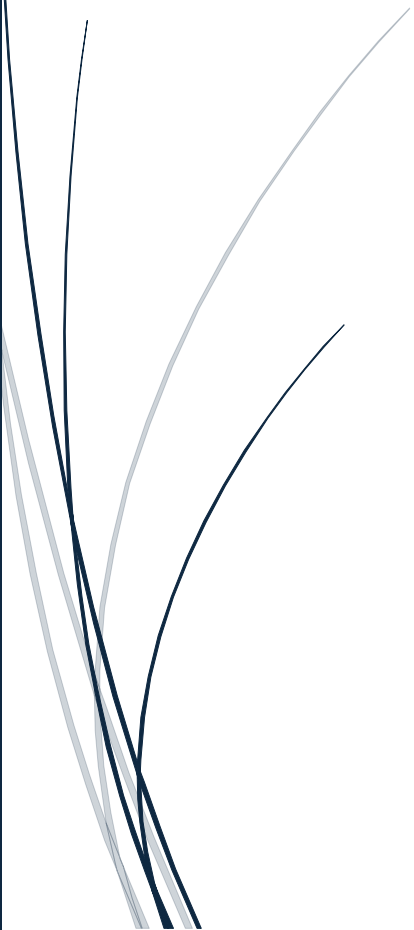




Multiple Regression Model

Dummy Independent Variables

TEAM 15 – Madhya Pradesh [23]



Ackshaiyaa V
EE/2023-25/001
Dilipkumaar N
EE/2023-25/008
Harish Chander P
EE/2023-25/011
Vasanth D
EE/2023-25/026

Computer Assignment – 2B: Dummy Independent Variables

Returns to Knowledge of English Language

This analysis is based on the second wave (2011-12) of the Indian Human Development Survey Data. For this study we are excluding the state of Madhya Pradesh with STATEID 23. Furthermore, we are taking in few conditions by keeping values where the age is between 15 to 65, $\text{indus_grp} = 9 \text{ \& } 999$, $\text{dwg} = 1$.

Total hours worked in the last one year is given by the variable *wrkhr* which has zero values in it, therefore we exclude a total of 31 observations with zero workhours, as the individuals with zero workhours cannot have a positive wage.

After giving the following conditions and creating necessary variables, the original dataset which had 204569 obs. & 61 variables, is now with a total of 62 variables & 47574 obs.

List of Dummy Variables used

English Knowledge:

$\text{DEK}_1 = \text{None}$

$\text{DEK}_2 = \text{Little}$

$\text{DEK}_3 = \text{Fluent}$

Gender:

$\text{DG}_1 = \text{Male}$

$\text{DG}_2 = \text{Female}$

1. Regress log of hourly wages on
 - i. Model 1: dummy variables based on educd;
 - ii. Model 2: eduyrs and squared eduyrs (sqeduyrs).

Question: Compare the goodness of fit in models 1 and 2 above and discuss the method you will use followed by the conclusion as to which model fits the data better? Are all the coefficients in the respective model statistically significant? Interpret the results of education in model 1 and discuss this in comparison to the interpretation in model 2.

Educational achievement of the workers which has four categories in this data set:

- i. None
- ii. Primary;
- iii. Middle;
- iv. secondary;
- v. Higher secondary;
- vi. Post Higher Secondary;

A regression between $\ln(\text{hrlywage})$ and education requires 5 dummy variables defined as follows:

DE2 = 1 if Educ.= primary, DE2 = 0 otherwise

DE3 = 1 if Educ.= Middle, DE3 = 0 otherwise

DE4 = 1 if Educ.= Secondary, DE4 = 0 otherwise

DE5 = 1 if Educ.= Higher Secondary, DE4 = 0 otherwise

DE6 = 1 if Educ.= Post Higher Secondary, DE4 = 0 otherwise

1. Modell_1:

$$\ln(\widehat{\text{hrlywage}}) = \alpha_0 + \alpha_1(DE_2)_i + \alpha_2(DE_3)_i + \alpha_3(DE_4)_i + \alpha_4(DE_5)_i + \alpha_5(DE_6)_i + u_i$$

$i=1,2,\dots,n$

$$\ln(\widehat{\text{hrlywage}}) = 2.452659 + 0.110565(DE_2)_i + 0.280693(DE_3)_i + 0.508918(DE_4)_i + 0.691570(DE_5)_i + 1.350539(DE_6)_i$$

Modell_2:

$$\ln(\widehat{\text{hrlywage}}) = \beta_0 + \beta_1 \text{eduyrs} + \beta_2 \text{eduyrs}^2$$

$$\ln(\widehat{\text{hrlywage}}) = 2.4707678 - 0.0046730 \text{eduyrs} + 0.0059541 \text{eduyrs}^2$$

The goodness of fit (Adjusted R^2) of model1_1 explains 17.52% of variations in $\ln(\text{hrlywage})$ whereas, model1_2 explains 18.16% with over 1% more explanation than model1_1.

We can use the **F-statistic** to infer which is a better model, considering only a small difference on the better fit for $\ln(\text{hrlywage})$:

$$F_{\text{Model1}_1} = 2022$$

$$F_{\text{model1}_2} = 5278$$

Therefore, for $F_{\text{tab}} = 2.605$ we are having a F_{model1_2} , having higher evidence than F_{Model1_1} against rejecting the null hypothesis of having no effect. Thua, we can say that Model1_2 is a better fit than model1_2.

In model1_1 all the coefficient estimates are statistically significant, which means that they all explain for the variations in $\ln(\text{hrlywage})$, whereas in model1_2 the coefficient estimate of education years is statistically insignificant which does not explain for the positive variations in $\ln(\text{hrlywage})$, therefore, as the education years keeps on increasing, the $\ln(\text{hrlywage})$ decreases.

TABLE 1: (Model1_1 & Model1_2)

		t-statistic	R^2
Model1_1	Primary	7.397	0.1752
	Middle	26.075	
	Secondary education	43.352	
	Higher secondary	44.233	
	Post higher secondary	93.856	
Model1_2	Education years(eduysrs)	-1.994	0.1816
	Square of education years(sqeduysrs)	35.985	

The tabulated values (Critical values) for both the models are the same with **1.644886**. For, $H_0: \beta_j = 0$; $H_1: \beta_j > 0$ to test the statistical significance of the coefficients. Therefore, using the values derived which is in the table, we can conclude that, for a one-sided t – test, except for the coefficient **β_2 of Education years**, all the coefficients are statistically significant with $t_{\text{cal}} > t_{\text{tab}}$.

2. Estimate three different models with log of hourly wages as dependent variable and the following three different models:
 - i. two dummy variables for English knowledge and excluding the dummy variable for 'none' which is the reference (or the omitted) group and including the intercept (constant term).
 - ii. three dummy variables for each of the three English knowledge dummies (that is without excluding the reference group of 'none') and including the intercept (constant term).
 - iii. three dummy variables for each of the three English knowledge dummies (that is without excluding the reference group of 'none') and excluding the intercept (constant term).

Question: Interpret the coefficients in models (1) and (3) and explain why model (2) does not give estimates for all the three dummy variables for English knowledge.

Level of English knowledge has 3 categories:

- i. None
- ii. Little
- iii. Fluent

A regression between $\ln(\text{hrlywage})$ and level of English Knowledge requires 2 dummy variables defined as follows:

$DEK_1 = 1$ if $ED3 = \text{None}$, $DEK_1 = 0$ (Baseline Category)

$DEK_2 = 1$ if $ED3 = \text{Little}$, $DEK_2 = 0$ otherwise

$DEK_3 = 1$ if $ED3 = \text{Fluent}$, $DEK_3 = 0$ otherwise

Model2_1:

$$\ln(\widehat{\text{hrlywage}}) = \alpha_0 + \alpha_1(DEK_2)_i + \alpha_2(DEK_3)_i + u_i \quad i=1,2,\dots,n$$

$$\ln(\widehat{\text{hrlywage}}) = 2.615063 + 0.569673(DEK_2)_i + 1.345528(DEK_3)_i$$

Model2_2:

$$\ln(\widehat{\text{hrlywage}}) = \alpha_0 + \alpha_1(DEK_1)_i + \alpha_2(DEK_2)_i + \alpha_3(DEK_3)_i + u_i \quad i=1,2,\dots,n$$

Model2_2 does not have an estimate equation.

Model3:

$$\ln(\widehat{\text{hrlywage}}) = \alpha_1(DEK_1)_i + \alpha_2(DEK_2)_i + \alpha_3(DEK_3)_i + u_i$$

$$\ln(\widehat{\text{hrlywage}}) = 2.615063(DEK_1)_i + 3.184736(DEK_2)_i + 3.960591(DEK_3)_i \quad i=1,2,\dots,n$$

With the presence of an intercept in model 2, the reference (or base) group which is the 'none' English Knowledge is explicitly explained for the average of $\ln(\text{hrlywage})$:

$$\text{Average} \left[\frac{\ln(\text{hrlywage})}{(\text{DEK})_i} \right] = \alpha_0 + \alpha_1 (\text{DEK})_i$$

Whereas, in the model 3, with no intercept in the model, such that the reference group (none) is now observed in the 'Little' and 'Fluent' dummy variable for the $\ln(\text{hrlywage})$ differences relative to the absence of the baseline (none).

Model2_2 does not give estimates for all the three dummy variables for English knowledge because we encounter a '**Dummy Variable Trap**'. A Dummy Variable Trap occurs when one or more dummy variables are redundant or multicollinear, which means they can be predicted from other variables. Here the intercept term gives estimates for the baseline category which we have included alongside all the 3 categories available in the English Knowledge variable 'ED3' which are 'None', 'Little' and 'Fluent', leading to dummy variable trap due to multicollinearity and leads in giving inaccurate estimates.

3. *Regress log of hourly wages on the following set of regressors. The regressors have to be incrementally added in the order mentioned below. Each number below will be one separate model*
 - i. *English knowledge- 'none' as the reference group;*
 - ii. *education in years (eduyrs) and its square (sqeduyrs);*
 - iii. *age (age) and squared age (sqage)*
 - iv. *gender- 'males' as the reference group*
 - v. *Interaction between gender and all other variables.*

Question:

- a. *Discuss how the coefficients of English knowledge change as you include additional variables in the models from (1) to (3) and what is the final conclusion about the role of English knowledge in addition to education and experience in explaining variation in hourly wages. The discussion should be in terms of the change in significance if any and changes in magnitude of the English knowledge variables.*
- b. *Discuss what is the difference between models (4) and (5)?*
- c. *If model (5) is the unrestricted model and model (3) is the restricted model then what is the test of hypothesis intending to do? Write the test statistic along with its degrees of freedom in this case and the conclusion if model (5) fits the data better or model (3). Based on the results of the test of hypothesis, what does this inform us about gender differences in log hourly wages?*

Level of English knowledge has 3 categories:

- i. None
- ii. Little
- iii. Fluent

A regression between $\ln(\text{hrlywage})$ and level of English Knowledge requires 2 dummy variables defined as follows:

$DEK_1 = ED3 = \text{None}$ (Baseline Category)

$DEK_2 = 1$ if $ED3 = \text{Little}$, $DEK_2 = 0$ otherwise

$DEK_3 = 1$ if $ED3 = \text{Fluent}$, $DEK_3 = 0$ otherwise

Gender has 2 categories

$DG_1 = \text{Male}$ (Baseline Category)

$DG_2 = 1$ if gender = Female, $DG_2 = 0$ otherwise

Model 1:

$$\ln(\widehat{\text{hrlywage}}) = \alpha_0 + \alpha_1(DEK_2)_i + \alpha_2(DEK_3)_i + u_i$$

$$\ln(\widehat{\text{hrlywage}}) = 2.615063 + 0.569673(DEK_2)_i + 1.345528(DEK_3)_i$$

TABLE 2 (model3_1)

Coefficients	Estimated	Standard error	T statistic	Pr(> t)
intercept	2.615063	0.004654	561.92	0
ED3Little	0.569673	0.010058	56.64	0
ED3Fluent	1.345528	0.016159	83.27	0

47571 degrees of freedom; \bar{R}^2 : 0.1597

Model 2:

$$\ln(\widehat{\text{hrlywage}}) = \alpha_0 + \alpha_1(DEK_2)_i + \alpha_2(DEK_3)_i + \alpha_3\text{edyrs} + \alpha_4\text{edyrs}^2 + u_i$$

$$\ln(\widehat{\text{hrlywage}}) = 2.4619513 + 0.2267344(DEK_2)_i + 0.7112066(DEK_3)_i + 0.0119569\text{edyrs} + 0.0030149\text{edyrs}^2$$

TABLE 3 (model3_2)

Coefficients	Estimated	Standard error	T statistic	Pr(> t)
Intercept	2.4619513	0.0068421	359.822	0
ED3Little	0.2267344	0.0120158	18.870	0
ED3Fluent	0.7112066	0.0212375	33.488	0
Education	0.0119569	0.0024089	4.964	0
Education ²	0.0030149	0.0001856	16.248	0

47569 degrees of freedom; \bar{R}^2 :0.2008

Model 3:

$$\ln(\widehat{\text{hrlywage}}) = \alpha_0 + \alpha_1(DEK_2)_i + \alpha_2(DEK_3)_i + \alpha_3\text{eduyrs} + \alpha_4\text{eduyrs}^2 + \alpha_5\text{age} + \alpha_6\text{age}^2 + u_i$$

$$\ln(\widehat{\text{hrlywage}}) = 1.759 + 0.2148(DEK_2)_i + 0.6837(DEK_3)_i + 0.02869\text{eduyrs} + 0.002162\text{eduyrs}^2 + 0.2676\text{age} + 0.0002208\text{age}^2$$

TABLE 4 (model3_3)

Coefficients	Estimated	Standard error	T statistic	Pr(> t)
Intercept	1.759	0.03664	48.018	0
ED3Little	0.2148	0.01191	18.043	0
ED3Fluent	0.6837	0.02105	32.471	0
eduyrs	0.02869	0.002449	11.827	0
sqeduyrs	0.002162	0.0001859	11.629	0
age	0.2676	0.001931	13.858	0
sqage	0.0002208	0.00002446	-9.028	0

47567 degrees of freedom; \bar{R}^2 : 0.2163

a) Interpretation:

As $|t_{\text{cal}}| > 2$, for a large dataset, we can evidently infer that all the coefficients of the models (1) to (3) are statistically significant for the variations in the $\ln(\text{hrlywage})$ and also by the fact of low p-value, the H_0 is rejected and thus proving to have an effect in the $\ln(\text{hrlywage})$.

There is a reduction in the value of coefficients from models (1) to (3) because at higher levels of experience and age, it implies that irrespective of English knowledge, the average hourly wages increase with experience and age, the need for English knowledge comes down

Earlier in the models with only dummy variables and by adding more continuous variables, the adjusted R^2 value improves upon such that we can imply that the coefficient estimates of English knowledge have proved to have more explanation in $\ln(\text{hrlywage})$

Despite the decreasing coefficient values, the positive signs on 'Little' and 'Fluent' across all models indicate that better English skills are associated with higher hourly wages, even when accounting for other factors.

Conclusion:

By the given information from the tables, we can see that the explanation of the dummy variables in English Knowledge tends to go low more and more aggregation of continuous variables into the model:

With addition of education years and squared education years(model3_2): The explanation of 'Little' moves from about **5.6% to 2.2%** and 'Fluent' moves to **7.1% form 13.4%**, which we can also say that due to the inclusion of the continuous variables into the model, the model's Adjusted R^2 value improves to **20%** variations of $\ln(\text{hrlywage})$ from **15.97%** with about **4% improvement** of fit to the model.

Now, with addition of age and squared age into the model, the model's fit moves to **21%** variations in $\ln(\text{hrlywage})$ which comes with a decrease in the magnitude of 'Little' and 'Fluent' further to **2.1% and 6.8%** respectively.

Therefore, when other factors like age and education are taken into account the direct contribution of English proficiency in the explanation of salary difference is diminished. But English Knowledge however, is considered to be statistically significant in determining hourly wages, underscoring its importance in the labour market in addition to education and experience.

The ultimate finding indicates that although having knowledge of English can have a favourable impact on earnings. It is only one component of a larger collection of characteristics such as age, education and experience that all work together to create disparities in salaries.

b)

Model3_4:

$$\ln(\widehat{\text{hrlywage}}) = \alpha_0 + \alpha_1(DEK_2)_i + \alpha_2(DEK_3)_i + \alpha_3\text{eduyrs} + \alpha_4\text{eduyrs}^2 + \alpha_5\text{age} + \alpha_6\text{age}^2 + \alpha_7(DG_2)_i + u_i$$

$$\ln(\widehat{\text{hrlywage}}) = 1.955 + 0.2047(DEK_2)_i + 0.6875(DEK_3)_i + 0.001852\text{eduyrs} + 0.003640\text{eduyrs}^2 + 0.02977\text{age} - 0.0002706\text{age}^2 - 0.4486(DG_2)_i$$

TABLE 5 (model3_4)

Coefficients	Estimated	Standard error	T statistic	Pr(> T)
Intercept	1.955	0.03584	54.547	0
ED3Little	0.2047	0.01158	17.670	0
ED3Fluent	0.6875	0.02048	33.563	0
eduyrs	0.001852	0.002455	-0.754	0.451
sqeduyrs	0.003640	0.0001831	19.875	0
age	0.02977	0.001879	15.844	0
sqage	-0.0002706	0.00002381	-11.363	0
genderfemales	-0.4486	0.008638	-51.927	0

47566 degrees of freedom; \bar{R}^2 : 0.2583

Model3_5:

$$\ln(\widehat{\text{hrlywage}}) = \alpha_0 + \alpha_1(DEK_2)_i + \alpha_2(DEK_3)_i + \alpha_3\text{eduyrs} + \alpha_4\text{eduyrs}^2 + \alpha_5\text{age} + \alpha_6\text{age}^2 + \alpha_7(DG_2)_i + \alpha_8(DEK_2)_i \cdot (DG_2)_i + \alpha_9(DEK_3)_i \cdot (DG_2)_i + \alpha_{10}\text{eduyrs} \cdot (DG_2)_i + \alpha_{11}\text{eduyrs}^2 \cdot (DG_2)_i + \alpha_{12}\text{age} \cdot (DG_2)_i + \alpha_{13}\text{age}^2 \cdot (DG_2)_i + u_i$$

$$\ln(\widehat{\text{hrlywage}}) = 1.882 + 0.1932(DEK_2)_i + 0.6508(DEK_3)_i + 0.003387\text{eduyrs} + 0.003007\text{eduyrs}^2 + 0.03445\text{age} - 0.0003265\text{age}^2 - 0.2351(DG_2)_i + 0.06567(DEK_2)_i \cdot (DG_2)_i + 0.1395(DEK_3)_i \cdot (DG_2)_i - 0.02791\text{eduyrs} \cdot (DG_2)_i + 0.002880\text{eduyrs}^2 \cdot (DG_2)_i - 0.01274\text{age} \cdot (DG_2)_i + 0.0001525\text{age}^2 \cdot (DG_2)_i$$

TABLE 6 (model3_5)

Coefficients	Estimate	Standard error	T statistic	Pr(> t)
Intercept	1.882	0.04239	44.394	0
ED3Little	0.1932	0.01278	15.081	0
ED3Fluent	0.6508	0.02285	28.486	0
eduyrs	0.003387	0.002935	1.154	0.24849
sqeduyrs	0.003007	0.0002116	14.212	0
age	0.03445	0.002256	15.267	0
sqage	-0.0003265	0.00002871	-11.371	0
genderfemales	-0.2351	0.07854	-2.996	0.00273
ED3Little:genderfemales	0.06567	0.03008	2.183	0.02904
ED3Fluent:genderfemales	0.1395	0.05147	2.711	0.00671
eduyrs:genderfemales	-0.02791	0.005454	-5.117	0
sqeduyrs:genderfemales	0.002880	0.0004320	6.666	0
age:genderfemales	-0.01274	0.004082	-3.122	0.00180
sqage:genderfemales	0.0001525	0.00005133	2.970	0.00298

47650 degrees of freedom; \bar{R}^2 :0.2611

Interpretation:

Model3_4 makes a simpler assumption that genders affect salaries in the same way independent of other factors. Model3_5 offers a more comprehensive explanation where gender interacts between all the other variables. This can be crucial in certain situations where male and female can experience different effect from the same variables such education and English knowledge. According to model 5 men's income are strongly impacted over years of schooling than women. This can point to the necessity for laws that support women's access to education or deal with possible gender discrimination in work place.

Categorical Variables (little, fluent): While Model3_4 might have a single coefficient for "little language skills" or "fluent language skills", Model3_5 likely has separate coefficients for men and women with these skills (female little, female fluent). This allows the model to capture gender-specific differences in the wage impact of language skills.

c)

Unrestricted Model:

Model3_5:

$$\begin{aligned}\ln(\widehat{\text{hrlywage}}) = & \alpha_0 + \alpha_1(DEK_2)_i + \alpha_2(DEK_3)_i + \alpha_3\text{eduyrs} + \alpha_4\text{eduyrs}^2 + \alpha_5\text{age} \\ & + \alpha_6\text{age}^2 + \alpha_7(DG_2)_i + \alpha_8(DEK_2)_i \cdot (DG_2)_i + \alpha_9(DEK_3)_i \cdot (DG_2)_i \\ & + \alpha_{10}\text{eduyrs} \cdot (DG_2)_i + \alpha_{11}\text{eduyrs}^2 \cdot (DG_2)_i + \alpha_{12}\text{age} \cdot (DG_2)_i \\ & + \alpha_{13}\text{age}^2 \cdot (DG_2)_i + u_i\end{aligned}$$

Restricted Model:

Model3_3:

$$\ln(\widehat{\text{hrlywage}}) = \alpha_0 + \alpha_1(DEK_2)_i + \alpha_2(DEK_3)_i + \alpha_3\text{eduyrs} + \alpha_4\text{eduyrs}^2 + \alpha_5\text{age} + \alpha_6\text{age}^2 + u_i$$

$$\begin{aligned}\ln(\widehat{\text{hrlywage}}) = & 1.759 + 0.2148(DEK_2)_i + 0.6837(DEK_3)_i + 0.02869\text{eduyrs} \\ & + 0.002162\text{eduyrs}^2 + 0.2676\text{age} + 0.0002208\text{age}^2\end{aligned}$$

Testing of Hypothesis:

Null Hypothesis: $H_0: \alpha_7, \alpha_8, \alpha_9, \alpha_{10}, \alpha_{11}, \alpha_{12}, \alpha_{13} = 0$

Alternate Hypothesis: $H_1: \alpha_7, \alpha_8, \alpha_9, \alpha_{10}, \alpha_{11}, \alpha_{12}, \alpha_{13} \neq 0$

$$F_{\text{cal}} = \frac{RSS_R - RSS_{UR} / q}{RSS_{UR} / (n - k - 1)_{UR}}$$

Here,

q = number of restrictions/ numerator degrees of freedom = $df_r - df_{ur}$

$n-k-1$ is the denominator degrees of freedom from unrestricted model = df_{ur}

therefor, we have $q = 7$.

$$F_{7, 47560} = 580.64$$

$$F_{tab} = 2.009783$$

Interpretation:

We reject the null hypothesis (H_0) in favour of alternative hypothesis (H_1) since the F statistic is significantly greater than the crucial threshold. This indicates that when compared to model 3 model 5 which incorporates gender interaction factors offers a noticeably better fit to the data.

According to the analysis there is still a gender pay disparity that persists with women earning much less than men. Wages and education are positively connected although the effect is stronger as education level rises. This is due to the fact that, the education coefficient alone only accounts for the impact on wages at zero years of education. The $squduyrs$ term is significant and positive coefficient suggests that if the value of further education rises with worker experience.