TOPIC – EFFECT OF MARITAL STATUS ON EMPLOYMENT IN INDIAN CONTEXT

NAME – MAMINA PANDA

INSTRUCTOR – Prof Amit Ray

SEMESTER – 4th

COURSE – ADVANCE ECONOMETRICS (IE 504)

1. Introduction:

India has experienced significant social and economic transformations over time, which have affected the makeup of the labour force.
It is essential to comprehend the connection between marital status and labour force participation in order to understand workforce dynamics and encourage economic growth. According to the labour supply theory, marriage may be one of the barriers preventing women from entering the workforce. However, broader institutional and societal factors as well as personal choices have an impact on how married women participate in the labour force. The opportunities that married women have in the workforce are shaped by cultural norms, expectations, and structural barriers, which affects their capacity to manage work and family obligations. Gender inequality in labour force participation rates can be sustained by societies where married women are pressured to put family responsibilities ahead of professional growth. On the other hand, married women may have more opportunities to enter the workforce and advance their careers in societies with more equal gender norms and supportive policies. Policymakers, employers, and individuals must all be aware of how marriage affects women's labour force participation because it illuminates the complex dynamics between genders in the workplace and provides guidance for initiatives aimed at advancing gender equality and women's economic empowerment.

2. Data Sources and variables:

I have taken the data from NFHS, India. From NFHS, DHS (Demographic Health and Survey) 2015-2016 data has been used. The DHS Program has earned a worldwide reputation for collecting and collecting and disseminating accurate, nationally representative data on fertility, family planning, maternal and child health, gender, HIV/AIDS, malaria and nutrition. The variables which I am using are v013(age), v024(states), v025(residence), v106(education level), v130(religion), v131(caste), v137(children), v501(marital status), v174(Employment status), v190(wealth index), v005(weights) and v002(Household number). I have also taken pre calculated states sex ratio from NFHS and merged it in the DHS data. The reason behind merging is the requirement of relevant variable as all the variables were not present in an IR DHS data.

3. Econometrics Model Specification:

Cross-sectional data study using descriptive and analytical approach. Our sample size after data cleaning and merging becomes 121,534 observations.

$$Yi = \beta o + \beta_1 X_1 + \beta_2 X_2 + \cdots \ldots\ldots \beta_k X_k + \epsilon_i$$

Under CLRM, E(Y_i|X) = βX_i = probability of dropout = p_i

Thus, we are estimating a linear probability model.

So basically, the Linear Probability Model (LPM) is a simple regression model used in econometrics and statistics to model binary outcome variables. In the LPM, the dependent variable is binary, meaning it can take on only two values, typically coded as 0 and 1. The model assumes that the relationship between the independent variables and the probability of the dependent variable taking on the value of 1 is linear.

Mathematically, the LPM can be expressed as:

$$E(Y_i|X) = \beta_0 + \beta_1 \sum_1^3 age\_dummy + \beta_2 \sum_1^4 caste\_category + \beta_3 \sum_0^2 marriage\_status + \beta_4 \sum_0^2 Socio\_Economics\_status + \beta_5 \sum_1^3 Religion + \beta_6 \sum_1^4 location + \beta_7\, education\_level + \beta_8 \sum_1^2 residence + \beta_9 * children$$

Our response variable is "Employment_status" which takes two distinct values

Employment_status (EMP) = 0, if not in the labour force, i.e., person is unemployed

= 1, otherwise

Where age_dummy = 1 if age group contains 15-19 and 20-24

age_dummy = 2 if age group contains 25-29, 30-34 and 35-39

age_dummy = 3 if age group contains 40-44 and 45-49

caste_category = 1 if caste is general category

caste_category = 2 if caste is scheduled caste

caste_category = 3 if caste is scheduled tribe

caste_category = 4 if caste is unknown

marriage_status = 0 if marital status is unmarried

marriage_status = 1 if marital status is married

marriage_status = 2 if marital status is widowed, divorced and separated

Socio_Economic_status = 0 if wealth index is poorest and poorer

Socio_Economic_status = 1 if wealth index is middle class

Socio_Economic_status = 2 if wealth index is richer and richest

Religion = 1 if religion is Hindu

Religion = 2 if religion is Muslim

Religion = 3 if religion is all other religion and no religion

Location = 1 if state is in Northern part of India

Location = 2 if state is in Southern part of India

Location = 3 if state is in Eastern part of India

Location = 4 if state is in Western part of India

Education_level = 0 if highest education level is no education

Education_level = 1 if highest education level is primary

Education_level = 2 if highest education level is secondary

Education_level = 3 if highest education level is higher

Residence = 1 if type of residence is urban

Residence = 2 if type of residence is rural

And children are discrete variable ranging from 0 to 9.

Since there could be potential endogeneity issue, I will run LPM with IV.

Reasons for Endogeneity issue:

i. Self-selection into marriage based on employment status.

ii. Reverse causality because marital status can affect employment status but marriage decision can also be explained whether or not women decide to join the labour force.

If this is the case, the above model estimates could be biased.

In order to solve the problem of endogeneity, the instrumental variable estimation technique (IV) will be employed.

Instrumental Variable (IV):

Sex ratio can serve as an instrument for marital status. The rationale is that sex ratio is unlikely to be directly related to employment status but may affect marital status. For example, in areas with a skewed sex ratio, there might be different marriage dynamics compared to areas with a balanced sex ratio.

Choosing Between LPM and IV:

Using **Wu-Hausman** test which will tell us whether to use simple LPM or IV.

This test is a statistical test used to assess the presence of endogeneity in regression models, particularly in the context of panel data or instrumental variables (IV) regression models. Hausman tests can be used to compare OLS and IV models. Under the null hypothesis, the OLS assumptions are not violated. In this case, both OLS and IV yield consistent estimates, but OLS is more efficient.

Potential Problem with LPM:

It can produce predicted probabilities that fall outside the [0, 1] range. This violates the probability constraint of binary outcome variables, as probabilities should always be within this range. Predicted probabilities outside this range can lead to unrealistic predictions and difficulties in interpretation.

If results we will get does not have any endogeneity issue then I will run logistic regression which will give direct causality of marital status on employment.

Logistic regression is a statistical technique used to model the probability of a binary outcome based on one or more independent variables. Unlike linear regression, which is used for continuous outcomes, logistic regression predicts the probability of an event occurring (e.g., success or failure, presence or absence) by fitting the data to a logistic curve. This curve maps the linear combination of the independent variables to a probability value between 0 and 1, making logistic regression well-suited for classification tasks. The coefficients obtained from logistic regression represent the log-odds of the outcome, allowing for the interpretation of the effects of the independent variables on the likelihood of the event.

$$P(Y=1|X) = \frac{1}{1+e^{-a}},$$

where a = $\beta_0 + \beta_1 \sum_1^3 age\_dummy_. + \beta_2 \sum_1^4 caste\_category_. + \beta_3 \sum_0^2 marriage\_status_. + \beta_4 \sum_0^2 Socio\_Economics\_status_. + \beta_5 \sum_1^3 Religion + \beta_6 \sum_1^4 location + \beta_7\, education\_level_. + \beta_8 \sum_1^2 residence + \beta_9 * \textbf{children}$

4. Model for Testing Assumptions:

We will run the LPM regression for testing the assumptions.

reg EMP i.age_dummy i.location i.residence i.education_level i.Religion i.caste_category i.children i.Socio_Economic_status i.marriage_status

Source	SS	df	MS
Model	1551.42087	27	57.460032
Residual	20202.4133	121,020	.1669345
Total	21753.8341	121,047	.179713947

Number of obs	=	121,048
F(27, 121020)	=	344.21
Prob > F	=	0.0000
R-squared	=	0.0713
Adj R-squared	=	0.0711
Root MSE	=	.40858

EMP	Coefficient	Std. err.	t	P>|t|	[95% conf. interval]	
age_dummy						
2	.1341972	.0034098	39.36	0.000	.127514	.1408804
3	.150644	.0042709	35.27	0.000	.142273	.1590149
location						
2	.0868235	.0039031	22.24	0.000	.0791736	.0944735
3	-.0209362	.0033343	-6.28	0.000	-.0274714	-.0144011
4	.071448	.003296	21.68	0.000	.064988	.0779081
residence						
2. rural	.0122152	.0029146	4.19	0.000	.0065027	.0179278
education_level						
1. primary	-.0077071	.0040954	-1.88	0.060	-.015734	.0003197
2. secondary	-.0552556	.0033203	-16.64	0.000	-.0617633	-.048748
3. higher	.0209619	.0048064	4.36	0.000	.0115415	.0303823
Religion						
2	-.0576546	.0037319	-15.45	0.000	-.0649691	-.0503401
3	.0543499	.0043484	12.50	0.000	.0458272	.0628726
caste_category						
2	-.0076179	.0059374	-1.28	0.199	-.0192551	.0040194
3	.0562729	.0069831	8.06	0.000	.0425862	.0699596
4	.0027313	.0228739	0.12	0.905	-.0421012	.0475638
children						
1	-.0308471	.0030201	-10.21	0.000	-.0367664	-.0249279
2	-.0488964	.0039165	-12.48	0.000	-.0565726	-.0412201
3	-.076753	.0071028	-10.81	0.000	-.0906744	-.0628317
4	-.0603461	.0141886	-4.25	0.000	-.0881556	-.0325367
5	-.0934866	.0260374	-3.59	0.000	-.1445196	-.0424537
6	-.0140398	.0572577	-0.25	0.806	-.1262639	.0981843
7	-.1106959	.0786851	-1.41	0.159	-.2649174	.0435256
8	-.2417844	.2889802	-0.84	0.403	-.8081809	.3246121
9	.0430356	.1827548	0.24	0.814	-.3151608	.401232
Socio_Economic_status						
1	-.0312045	.0033368	-9.35	0.000	-.0377447	-.0246644
2	-.0967534	.0034167	-28.32	0.000	-.10345	-.0900567
marriage_status						
1	-.0627457	.0038069	-16.48	0.000	-.0702072	-.0552842
2	.1357373	.0069514	19.53	0.000	.1221126	.1493619
_cons	.2367085	.00792	29.89	0.000	.2211855	.2522315

Interpreting results:

Being married is associated with decrease in probability of employment by approximately 0.063, while being divorced, widowed, separated is associated with increase of 0.136.

From the F statistics (p-value < 5%) we can see that there is an overall significance of the model.

4.1. Testing Assumptions of LPM:

Assumption 1: Multicollinearity:

Variance inflation factor (VIF): VIF measures the strength of correlation between the explanatory variables in our model by regressing each explanatory variable on all the other explanatory variables. If VIF > 10, then the explanatory variable is strongly correlated with the other explanatory variable, and so is redundant.

Variable	VIF	1/VIF
age_dummy		
2	2.07	0.482858
3	2.23	0.448272
location		
2	1.33	0.754257
3	1.69	0.590267
4	1.55	0.646780
2.residence	1.31	0.765572
education_~l		
1	1.32	0.756754
2	2.00	0.500756
3	1.75	0.569883
Religion		
2	1.28	0.779253
3	1.43	0.699967
caste_cate~y		
2	4.05	0.246742
3	4.40	0.227429
4	1.07	0.938287
children		
1	1.18	0.845490
2	1.15	0.865834
3	1.05	0.948339
4	1.01	0.987054
5	1.01	0.994921
6	1.00	0.998830
7	1.00	0.998836
8	1.00	0.999510
9	1.00	0.999669
Socio_Econ~s		
1	1.35	0.738308
2	2.04	0.490722
marriage_s~s		
1	2.17	0.461885
2	1.40	0.712785
Mean VIF	1.62	

Here, Mean VIF= 1.62< 10, hence there is no presence of Multicollinearity. **Assumption 1 is satisfied.**

4.2. Assumption 2: Homoscedasticity:

Assumption 2 requires constant variance of the error term, i.e., residuals are distributed with equal variance at each level of the predictor variable. However, violation can occur as heteroskedasticity when the residuals are not distributed with equal variance. This unequal scatteredness indicates a systematic change in the spread of the residuals over the range of measured values. So, the estimator will still be unbiased and linear but it will no longer be efficient.

To detect heteroscedasticity, we use the following method:

Breusch-Pagan test for heteroskedasticity:

- Null Hypothesis, H_o: Homoscedasticity is present in our model
- Alternative Hypothesis, H_A: Heteroscedasticity is present in our model

```
Breusch-Pagan/Cook-Weisberg test for heteroskedasticity
Assumption: Normal error terms
Variable: Fitted values of EMP

H0: Constant variance

    chi2(1) = 5276.74
Prob > chi2 =  0.0000
```

Since p value is less than 0.05 hence, we will reject the null hypothesis which means that there is heteroscedasticity in our model.

To correct for the heteroscedasticity, we run robust with our regression equation to get coefficients standard error adjusted for heteroscedasticity.

4.3. Assumption 3: Normality:

The error term must be normally distributed. Here, we will use Jarque Bera test for Normality.

```
. jb residuals
Jarque-Bera normality test:  2.6e+04 Chi(2)      0
Jarque-Bera test for Ho: normality:
```

Since p value is less than 0.05 hence, we will reject the null hypothesis that error terms are normally distributed.

4.4. Assumption 4: Linearity in parameter:

$$\mathbf{E}(Y_i|\mathbf{X}) = \boldsymbol{\beta_0} + \boldsymbol{\beta_1} \sum_1^3 \boldsymbol{age\_dummy} + \boldsymbol{\beta_2} \sum_1^4 \boldsymbol{caste\_category} + \boldsymbol{\beta_3} \sum_0^2 \boldsymbol{marriage\_status} + \boldsymbol{\beta_4} \sum_0^2 \boldsymbol{Socio\_Economics\_status} + \boldsymbol{\beta_5} \sum_1^3 \boldsymbol{Religion} + \boldsymbol{\beta_6} \sum_1^4 \boldsymbol{location} + \boldsymbol{\beta_7}\, \boldsymbol{education\_level} + \boldsymbol{\beta_8} \sum_1^2 \boldsymbol{residence} + \boldsymbol{\beta_9} * \mathbf{children}$$

We can check that all the parameters in the above LPM equation is linear, hence assumption 4 is satisfied.

4.5. Assumption 5: Checking for Autocorrelation

The Durbin-Watson test is a statistical test used to detect autocorrelation in the residuals of a regression model. Autocorrelation occurs when the residuals of a regression model are correlated with each other, indicating that there is some pattern or structure in the data that the model has not captured.

The Durbin-Watson test statistic ranges from 0 to 4. A value around 2 indicates no autocorrelation, while values significantly lower than 2 suggest positive autocorrelation, and values significantly higher than 2 suggest negative autocorrelation.

```
. dwstat

Number of gaps in sample = 84

Durbin-Watson d-statistic( 20,121048) =  1.703696

.
```

In my case the d-statistics is coming out to be 1.7 which is very close to 2, hence there is no autocorrelation.

4.6. Assumption 5: Checking for Endogeneity:

Explanatory variables need to be exogenous, i.e., determined by factors outside the model, for the estimator to be unbiased. Violation leads to endogeneity creating bias when the explanatory variables are correlated with the error term in the regression,

```
Tests of endogeneity
H0: Variables are exogenous

Durbin (score) chi2(1)          =  .000955  (p = 0.9754)
Wu-Hausman F(1,121028)          =  .000954  (p = 0.9754)
```

Here the p value is greater than 0.05 we will accept the null hypothesis which concludes that there is no endogeneity issue in our model.

Since there is no endogeneity, we can now run a binary logistic regression to get better picture instead of IV:

- As Logistic regression models the probability of a binary outcome using the logistic function, which constrains the predicted probabilities to fall between 0 and 1. In contrast, the LPM does not impose such constraints, leading to predicted probabilities that may fall outside the valid probability range.
- Logistic regression does not require the assumption of homoscedasticity (constant variance of errors) that is necessary for valid inference in the linear probability model. This makes logistic regression more robust in the presence of heteroscedasticity.
- In large samples, logistic regression estimates are more efficient and have smaller standard errors compared to the linear probability model. This is because logistic regression estimates are based on maximum likelihood estimation, which is asymptotically efficient.

5. Estimating Logistic Regression Equation:

$$\text{Logit}\,(P_i) = \beta_0 + \beta_1 \sum_1^3 age\_dummy_{.} + \beta_2 \sum_1^4 caste\_category_{.} + \beta_3 \sum_0^2 marriage\_status_{.} + \beta_4 \sum_0^2 Socio\_Economics\_status_{.} + \beta_5 \sum_1^3 Religion + \beta_6 \sum_1^4 location + \beta_7\, education\_level_{.} + \beta_8 \sum_1^2 residence + \beta_9 * children$$

Regressing the above equation:

Logistic regression — Number of obs = 121,048; LR chi2(19) = 8811.96; Prob > chi2 = 0.0000

Log likelihood = -61578.3 — Pseudo R2 = 0.0668

EMP	Odds ratio	Std. err.	z	P>|z|	[95% conf.	interval]
age_dummy						
2	2.491323	.0569889	39.90	0.000	2.382094	2.605561
3	2.721233	.0733723	37.13	0.000	2.58116	2.868908
location						
2	1.667175	.0386087	22.07	0.000	1.593195	1.74459
3	.8876628	.018681	-5.66	0.000	.8517935	.9250426
4	1.52549	.0305117	21.11	0.000	1.466846	1.58648
residence						
2. rural	1.072388	.0192751	3.89	0.000	1.035267	1.11084
education_level						
1. primary	.973454	.0224349	-1.17	0.243	.9304606	1.018434
2. secondary	.7355752	.0143296	-15.76	0.000	.7080191	.7642037
3. higher	1.182476	.0334435	5.93	0.000	1.118711	1.249874
Religion						
2	.6673852	.016802	-16.06	0.000	.6352531	.7011426
3	1.386878	.0341897	13.27	0.000	1.32146	1.455533

caste_category						
2	.9550521	.0371951	-1.18	0.238	.8848639	1.030808
3	1.324257	.057902	6.42	0.000	1.215498	1.442748
4	1.012423	.1465192	0.09	0.932	.7623864	1.344464
children	.8605197	.0077122	-16.76	0.000	.8455361	.8757687
Socio_Economic_st~s						
1	.833421	.0162438	-9.35	0.000	.8021842	.8658742
2	.5526882	.0115839	-28.29	0.000	.5304442	.5758651
marriage_status						
1	.6359652	.0156409	-18.40	0.000	.6060367	.6673716
2	1.570834	.0598949	11.84	0.000	1.457721	1.692723
_cons	.2781736	.0139941	-25.43	0.000	.2520546	.3069992

Note: **_cons** estimates baseline odds.

Note:

$$\text{Odds}=\frac{\text{the probability of an event favourable to an outcome}}{\text{probability of an event against the same outcome}}$$

Interpreting the results of logistic regression:

Probability is constrained between zero and one and odds are constrained between zero and infinity. And odds ratio is the ratio between odds.

Individuals in caste categories 2(SC) and 3(ST) have higher odds of employment compared to the reference category which is General category, with odds ratios of approximately 0.96 and 1.32, respectively.

Caste category 4(Unknown) shows a negligible effect on employment compared to the reference category.

Each additional child is associated with a decrease in the odds of employment by approximately 0.86 times, holding all other variables constant.

Individuals residing in rural areas (category 2) have approximately 1.07 times the odds of employment compared to the urban areas, with a 95% confidence interval.

Individuals in marriage status category 1(married) have approximately 0.64 times the odds of employment compared to the reference category(unmarried), while those in category 2 (widowed, divorced and separated) have approximately 1.57 times the odds.

5.1. Testing Assumptions for Logistic Regression:

5.1.1. Assumption 1: Binary nature of Outcome variable

Outcome variable should be binary in nature, which is true in this model. Thus, the **Assumption 1** is satisfied.

5.1.2. Assumption 2: Multicollinearity:

	age_du~y	location	reside~e	educat~l	Religion	caste_~y	children
age_dummy	1.0000						
location	0.0071	1.0000					
	0.0139						
residence	-0.0240	0.0006	1.0000				
	0.0000	0.8266					
education_~l	-0.3466	-0.0629	-0.2258	1.0000			
	0.0000	0.0000	0.0000				
Religion	0.0014	-0.0764	-0.0524	0.0395	1.0000		
	0.6374	0.0000	0.0000	0.0000			
caste_cate~y	0.0082	0.2410	0.0611	-0.0474	0.2225	1.0000	
	0.0045	0.0000	0.0000	0.0000	0.0000		
children	-0.0947	0.0061	0.0806	-0.0708	0.0081	0.0268	1.0000
	0.0000	0.0327	0.0000	0.0000	0.0049	0.0000	
Socio_Econ~s	0.0384	-0.1397	-0.4503	0.4294	0.0721	-0.1367	-0.0973
	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
marriage_s~s	0.5667	0.0466	0.0320	-0.3209	-0.0480	0.0148	0.1602
	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

	Socio_~s	marria~s
Socio_Econ~s	1.0000	
marriage_s~s	-0.0482	1.0000
	0.0000	

Results:

Each cell in the table represents the correlation coefficient between two variables. For example, the correlation coefficient between age_dummy and location is 0.0071, indicating a very weak positive correlation.

Similarly, the correlation coefficient between education level and caste category is -0.3466, indicating a moderate negative correlation.

Second, table provides correlation coefficients between different pairs of variables, similar to the first table. For example, the correlation coefficient between marriage_status and Socio_Economic_status is -0.0482, indicating a very weak negative correlation.

Thus, there is no perfect multicollinearity in the model, **Assumption 2 is satisfied.**

6. Computing marginal effects of Logistic Regression equation:

```
Average marginal effects                          Number of obs = 121,048
Model VCE: OIM

Expression: Pr(EMP), predict()
dy/dx wrt:  2.age_dummy 3.age_dummy 2.location 3.location 4.location 2.residence
            1.education_level 2.education_level 3.education_level 2.Religion
            3.Religion 2.caste_category 3.caste_category 4.caste_category children
            1.Socio_Economic_status 2.Socio_Economic_status 1.marriage_status
            2.marriage_status
```

	dy/dx	Delta-method std. err.	z	P>|z|	[95% conf.	interval]
age_dummy						
2	.1409878	.0032811	42.97	0.000	.1345569	.1474186
3	.1581579	.0042601	37.13	0.000	.1498084	.1665075
location						
2	.0892924	.0041608	21.46	0.000	.0811375	.0974473
3	-.0178916	.0031591	-5.66	0.000	-.0240834	-.0116997
4	.0723832	.0034148	21.20	0.000	.0656904	.0790761
residence						
2. rural	.0115643	.0029557	3.91	0.000	.0057712	.0173574
education_level						
1. primary	-.004743	.0040562	-1.17	0.242	-.0126931	.003207
2. secondary	-.0507483	.0032926	-15.41	0.000	-.0572016	-.044295
3. higher	.0307752	.0052375	5.88	0.000	.02051	.0410404
Religion						
2	-.0615794	.0035497	-17.35	0.000	-.0685368	-.0546221
3	.059157	.0046657	12.68	0.000	.0500125	.0683015
caste_category						
2	-.0075758	.006478	-1.17	0.242	-.0202726	.0051209
3	.0497806	.0075103	6.63	0.000	.0350606	.0645005
4	.0020621	.0242316	0.09	0.932	-.0454309	.0495551
children	-.0250052	.0014873	-16.81	0.000	-.0279204	-.0220901
Socio_Economic_st~s						
1	-.0331001	.0035167	-9.41	0.000	-.0399927	-.0262075
2	-.0981856	.0034465	-28.49	0.000	-.1049406	-.0914307
marriage_status						
1	-.078274	.0044146	-17.73	0.000	-.0869264	-.0696216
2	.0914731	.0079317	11.53	0.000	.0759273	.1070189

Note: dy/dx for factor levels is the discrete change from the base level.

Interpretation:

Individuals in location category 2 (South) have a marginal effect of approximately 0.089 on the probability of employment compared to the reference category (North), with a 95% confidence interval of [0.081, 0.097].

Location categories 3(East) and 4(west) show different effects on employment compared to the reference category, with marginal effects of approximately -0.018 and 0.072, respectively.

Each additional child is associated with a decrease in the probability of employment by approximately 0.025, holding all other variables constant.

Individuals in religion category 2 (Muslim) have a marginal effect of approximately -0.062 on the probability of employment compared to the reference category (Hindu), while those in category 3 (others) have a marginal effect of approximately 0.059

Individuals in caste categories 2(SC) and 3(ST) have marginal effects of approximately -0.008 and 0.050 on the probability of employment, respectively. Caste category 4(other) shows a negligible marginal effect on employment compared to the reference category (general category).

Individuals in socioeconomic status category 1(middle class) have a marginal effect of approximately -0.033 on the probability of employment compared to the reference category, while those in category 2(rich class) have a marginal effect of approximately -0.098.

Individuals in marriage status category 1 (married) have a marginal effect of approximately -0.078 on the probability of employment compared to the reference category, while those in category 2(widowed, divorced and separated) have a marginal effect of approximately 0.091.

7. Testing Goodness of fit of Model:

Using Hosmer-Lemeshow Test for Goodness of Fit:

The Hosmer-Lemeshow test is a goodness-of-fit test commonly used in logistic regression to assess how well the model fits the observed data. It evaluates whether the predicted probabilities from the logistic regression model match the observed outcomes.

```
Goodness-of-fit test after logistic model
Variable: EMP

      Number of observations =  121,048
Number of covariate patterns =    7,955
          Pearson chi2(7935) = 11101.01
                 Prob > chi2 =   0.0000
```

Here the p value is less than 0.05 hence our model is not a good fit.

8. Limitations of Analysis:

The independent variables may not be able to adequately predict or explain the variation in the dependent variable, as indicated by the low pseudo-R-squared (0.06). This could indicate that significant variables are missing from the model or that the model doesn't accurately represent the relationship between the variables. The majority of the literature analysing the connection between marital status and employment status uses more sophisticated models and complex model, which may place limitations on the methodology this project chooses.

9. Conclusion:

In this assignment I tried to examine the causality between marital status and employment status in India by using the data from NFHS 2015 – 2016. First through LPM I tried to test whether there is any endogeneity or not, after confirming that there is no endogeneity I run the logistic regression to get the casual effect, and it shows that married women have lower probability (-0.078) of being employed as compared to unmarried women and divorced and widowed have higher chance of being employed as compare to unmarried. Most of the coefficient are statistically significant but after testing the model through Hosmer-Lemeshow test we got know that logistic model is not appropriate for this and this could be for many reasons