Assignment 2

We are conducting the study for the state of West Bengal using IHDS-II data

Q 1a) We regress Y (expenditure share on food items) on X_2(log annual per capita income) and X_3(log of n- household members)

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + e$$

The estimated regression function can be written as

$$\hat{Y} = b_1 + b_2 X_2 + b_3 X_3$$

The results of the multiple regression can be summarised in the table given below

Multiple Linear regression – Y on X_2 and X_3

Food share	Coef.	St. Error	t-value	p-value	[95% Conf	Interval]	Sig
ln(apci)	-0.081(b_2)	0.003	-26.86	0.000	-0.087	-0.075	***
ln(hhsz)	-0.023(b_3)	0.007	-3.47	0.001	-0.036	-0.010	***
Constant	1.406	0.033	43.08	0.000	1.342	1.470	***

Mean dependent var	0.579	SD dependent var	0.157
R-squared	0.253	Number of observations	2142.000
F-test	361.556	Prob > F	0.000
Akaike crit. (AIC)	-2476.185	Bayesian crit. (BIC)	-2459.177

*** $p<0.01$, ** $p<0.05$, * $p<0.1$

$\hat{Y}$ = 1.406 - 0.081*ln(apci) - 0.023*ln(hhsz) is our estimated regression line.

s.e = (0.033) (0.003) (0.007)

t = (43.08) (-26.86) (-3.47) d.f = 2139

(0.000)* (0.000)* (0.001)* r^2 = 0.253

Now, we perform a series of simple regressions:

1. Regress Y on ln(apci) and obtain the residuals e1_2

Linear regression

food share	Coef.	St Error	t-value	p-value	[95% Conf	Interval]	Sig
ln(apci)	-0.079	0.003	-26.60	0.000	-0.085	-0.073	***
Constant	1.355	0.029	46.23	0.000	1.298	1.413	***

Mean dependent var	0.579	SD dependent var	0.157
R-squared	0.248	Number of observations	2142.000
F-test	707.395	Prob > F	0.000
Akaike crit. (AIC)	-2466.142	Bayesian crit. (BIC)	-2454.803

*** $p<0.01$, ** $p<0.05$, * $p<0.1$

$$\hat{Y} = 1.355 - 0.079*\ln(apci)$$

2. Regress ln(hhsz) on ln(apci) on and obtain the residuals e3_2

Linear regression

ln(hhsz)	Coef.	St Error	t-value	p-value	[95% Conf	Interval]	Sig
ln(apci)	-0.078	0.009	-8.48	0.000	-0.096	-0.060	***
Constant	2.137	0.090	23.69	0.000	1.960	2.314	***

Mean dependent var	1.375	SD dependent var	0.449
R-squared	0.029	Number of observations	2418.000
F-test	72.003	Prob > F	0.000
Akaike crit. (AIC)	2917.466	Bayesian crit. (BIC)	2929.048

*** p<0.01, ** p<0.05, * p<0.1

ln(hhsz) = 2.137 – 0.078*ln(apci)

3. Regress e3_2 on e1_2 and the obtain the slope coefficient $b_{13.2}$ to check if it is equal to b_3

Linear regression – e1_2 on e3_2

e1_2	Coef.	St Error.	t-value	p-value	[95% Conf	Interval]	Sig
e3_2	-0.023($b_{13.2}$)	0.007	-3.47	0.001	-0.036	-0.010	***
Constant	0.000	0.003	0.05	0.963	-0.006	0.006	

Mean dependent var	0.000	SD dependent var	0.136
R-squared	0.006	Number of observations	2142.000
F-test	12.066	Prob > F	0.001
Akaike crit. (AIC)	-2478.185	Bayesian crit. (BIC)	-2466.846

*** p<0.01, ** p<0.05, * p<0.1

e1_2 = 0 – 0.023*e3_2 is the regression line of error regression.

s.e = (0.003) (0.05)
t = (-3.47) (-0.05) d.f = 2140

We can see that the slope coefficient in the regression of e1_2 on e3_2 is $b_{13.2}$ = -0.023 which is equal to b_2 in the multiple regression model.
So, $b_3 = b_{13.2}$ (proved)

Q 1 b) Residual Analysis: Here, we have used the rvf plot to plot the residuals e1_2, e3_2 and e1_23 against the corresponding fitted values.

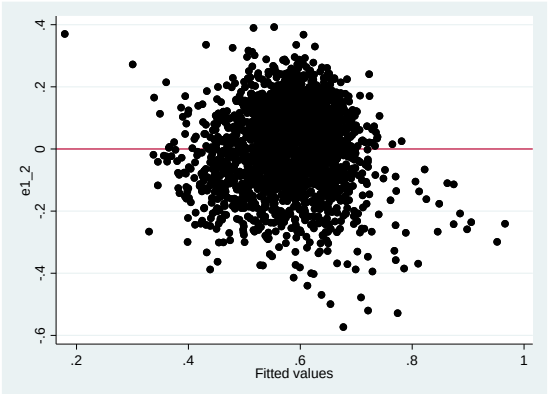

Figure 1: e1_2

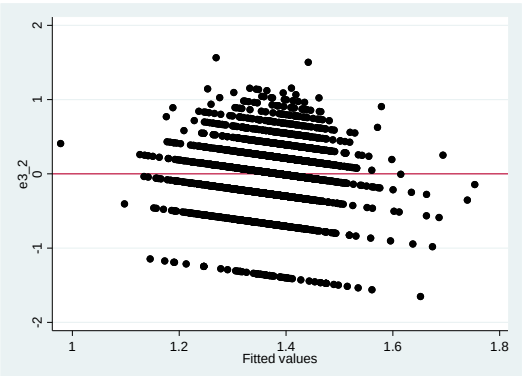

Figure 2: e3_2

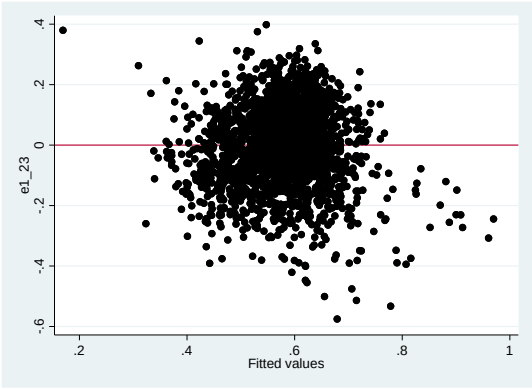

Figure 3: e1_23

Econometric intuition behind the rvf plot: The Residual versus fitted plot is a useful tool to find out how good our model fits the data and also to catch heteroscedasticity (non-constant variance of error terms). If the error terms show some specific pattern like increasing with rise in X-values, then we can say that our error term is showing heteroscedasticity or we can even have a case like the error terms are showing some non-linear pattern.
But if the error terms are scattered evenly, then it is homoscedastic.

If we look at the above residual plots, it is quite clear that there is no specific pattern in each one of them and the errors are more or less evenly distributed.
Now, let us compare figure 1 (only one explanatory variable – ln (per capita income)) and figure 3 (multiple regression), we can see that the rvf plots are more or less similar which implies that adding ln (household size) may not be that insightful in explaining food share.

Q1b (i) **Testing if the mean of the residual terms is 0:** It is very important to test if the mean of the residual terms is 0 to understand if our model follows the CLRM assumptions.

(a) H_0: e1_2 = 0

One-sample t test for e1_2

	Obsn	Mean	St Error	t-value	p-value
e1 2	2418	0	.003	0	1

Ha: mean < 0 Ha: mean≠ 0 Ha: mean > 0

P(T < t) = 0.5000 P(|T| > |t|) = 1.0000 P(T > t) = 0.5000

Conclusion - As p-value is greater than 0.01, we fail to reject the null hypothesis.

(b) H_0: e3_2 = 0

One - sample t test for e3_2

	Obsn	Mean	St Error	t-value	p-value
e3 2	2418	0	.009	0	1

Conclusion - As p-value is greater than 0.01, we fail to reject the null hypothesis

(c) H_0: e1_23 = 0

One-sample t test for e1_23

	Obsn	Mean	St Error	t-value	p-value
e1 23	2418	0	.003	0	1

Conclusion - As p-value is greater than 0.01, we fail to reject the null hypothesis Finally, we can say that the mean of the error terms in our fitted models is 0 according to our sample.

Q1b (ii) **Testing the normality of the residual terms:** The sk-test has been used here to test the normality of the residual terms- e1_2, e2_3, e1_23. The results have been summarised in the table given below.

Skewness-kurtosis test for Normality of residuals

				------ joint ------	
Residuals	Observations	P(Skewness)	P(Kurtosis)	Adj-Chi Sq.(2)	Prob >Chi (2)
e1_2	2,418	0.000	0.056	45.230	0.000
e2_3	2,418	0.000	0.000	.	0.000
e1_23	2,418	0.000	0.032	52.890	0.000

Interpretation: We can reject the hypothesis that the error terms are normally distributed as all of the p-values are 0 which indicates that it is significantly different from a Normal distribution with a combined measure of skewness-kurtosis as indicated by the last column. However, on the basis of kurtosis alone we fail to reject that e1_2 is normally distributed at 5 % level and e1_23 at 1 % level. The skewness measure alone also rejects the null hypothesis of normality. Note: If we look at the adj-Chi Square column for e2_3, it is indicated by a '.' indicating an absurdly large Chi-Square, the data are most certainly not Normal.

Question 2. We have 8 variables on demographic composition:

NADULTM_n, NADULTF_n, NCHILDM, NCHILDF, NTEENM, NTEENF, NELDERM and NELDERF
NPERSONS is the total number of households as given in the database.

The two new variables generated are NADULTM_n and NADULTF_n which stands for adult males and females above age 21 but less than 60.

NADULTM_n = NADULTM – NELDERM

NADULTF_n = NADULTF – NELDERF

We have to check whether NPERSONS is equal to sum of all the demographic variables.

So, we generate a new variable nn = sum of all the demographic variables.

We also generate aa = 1 if nn = NPERSONS and 0 otherwise.

Tabulation of aa

	Freq.	Percent	Cum.
0	2	0.08	0.08
1	2433	99.92	100.00

But, as we are working with West Bengal data, there are two households for which nn is not equal to NPERSONS. In both households, there is a grandchild whose age is not mentioned (age is missing). It might be because 1) baby is just born or 2) Age is unknown. Hence this individual is not there in NCHILDF and NCHILDM and hence total addition will not be matched with NPERSONS. So, we modify accordingly.

Tabulation of aa

	Freq.	Percent	Cum.
1	2435	100.00	100.00

So, we see that after taking into account the two children, nn = NPERSONS.

Now, we have generated 8 new variables, each denoting the proportion in each of the following age-groups with their gender

1. NCHILDF/NPERSONS = n_1
2. NCHILDM/NPERSONS = n_2
3. NADULTM/NPERSONS = n_3
4. NADULTF/NPERSONS = n_4
5. NTEENM/NPERSONS = n_5
6. NTEENF/NPERSONS = n_6
7. NELDERM/NPERSONS = n_7
8. NELDERF/NPERSONS = n_8

Q2 a) Firstly, we have tried to regress foods-share(Y) on all of the variables generated till now i.e., X_2, X_3, n_1, n_2, n_3, n_4, n_5, n_6, n_7 and n_8

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \sum_{i=1-8} \alpha(n_i) + e_i$$

But, while doing this in STATA, it shows an error due to presence of multicollinearity as $\sum_{i=1-8}(n_i) = 1$ (a perfectly linear relationship between the variables)

Linear Multiple Regression

Food-share	Coef.	St Error	t-value	p-value	[95% Conf	Interval]	Sig
X_2	-0.079	0.003	-24.74	0.000	-0.086	-0.073	***
X_3	-0.038	0.008	-4.87	0.000	-0.053	-0.022	***
n1	0.077	0.032	2.39	0.017	0.014	0.141	**
n2	0.063	0.033	1.94	0.052	-0.001	0.128	*
n3	0.070	0.033	2.11	0.035	0.005	0.135	**
n4	-0.029	0.032	-0.93	0.354	-0.092	0.033	
n5	0.028	0.037	0.76	0.450	-0.044	0.099	
n6	0.000	.	.	.	.	.	
n7	-0.064	0.025	-2.59	0.010	-0.113	-0.016	**
n8	0.033	0.024	1.38	0.168	-0.014	0.080	
Constant	1.381	0.041	33.36	0.000	1.299	1.462	***

Mean dependent var	0.579	SD dependent var	0.157
R-squared	0.262	Number of observations	2142.000
F-test	84.042	Prob > F	0.000
Akaike crit. (AIC)	-2488.774	Bayesian crit. (BIC)	-2432.079

*** $p<0.01$, ** $p<0.05$, * $p<0.1$

Here, as we can see, STATA drops n_6 due to collinearity.

The estimated model is $\hat{Y} = 1.381 - 0.079 X_2 - 0.038 X_3 + 0.077 n1 + 0.063 n2 + 0.070 n3 - 0.029 n4 + 0.028 n5 - 0.064 n7 + 0.033 n8$

Q 2 b) The coefficient of X_3(ln(hhsz)) is -0.038. In the case of only two variables as in question 1, it was -0.023. So, the sign is still negative and if we look at the p-value, it is 0<0.01 which implies that log (household size) is statistically significant at 1% level of significance. The magnitude has increased in absolute value. The interpretation of the coefficient of X_3 is that:

As household size increases by 1% keeping other factors constant, food share decreases, on an average, by 3.8%.

Q 2 c) If we look carefully at the coefficients of n1, n2, n3, n4, n5, n6, n7, n8 , we observe that the coefficient of n_1 and n_3 is higher which stands for the proportions of child females and elder males which is quite natural as adult males tend to have relatively more impact on food share due to their labor and also structure i.e., they are generally taller and have more weight. The coefficient of n_3 is also statistically significant.

However, coefficients of n_4,n_5,n_8 which stands for adult females, teen males and elder females are statistically insignificant which implies they have relatively less impact on food share. The statistical insignificance of teen males is surprising as they do have an impact on food share due to their growing years. The results can also highlight there is a discrimination against females in food consumption in West Bengal.

Q 3) **Re-estimation of model using dummy variables**:

HQ4 2.0 N in household	Freq.	Percent	Cum.
1	63	2.59	2.59
2	266	10.92	13.51
3	503	20.66	34.17
4	654	26.86	61.03
5	435	17.86	78.89
6	241	9.90	88.79
7	126	5.17	93.96
8	67	2.75	96.71
9	34	1.40	98.11
10	24	0.99	99.10
11	11	0.45	99.55
12	8	0.33	99.88
13	1	0.04	99.92
17	1	0.04	99.96
19	1	0.04	100.00

Looking at the table above, we see that majority of the household size lies between 1 to 7. So, we create 8 dummy variables, first 7 of them representing the household sizes from 1-7 and the last one representing household sizes > = 8. The last category is the reference category.

The results from regression are summarised in the table given below:

Linear regression – Using dummy variables for different household sizes

Food-share	Coef.	St. Error.	t-value	p-value	[95% Conf	Interval]	Sig
ln(apci)	-0.078	0.003	-24.34	0.000	-0.085	-0.072	***
n2	-0.011	0.027	-0.41	0.682	-0.064	0.042	
n3	-0.004	0.025	-0.17	0.866	-0.054	0.046	
n4	-0.117	0.028	-4.14	0.000	-0.173	-0.062	***
n5	-0.047	0.030	-1.60	0.110	-0.105	0.011	
n6	-0.080	0.032	-2.48	0.013	-0.144	-0.017	**
n7	-0.080	0.026	-3.11	0.002	-0.130	-0.029	***
n8	0.013	0.024	0.54	0.590	-0.035	0.061	
hhszgr1	0.117	0.024	4.89	0.000	0.070	0.163	***
hhszgr2	0.054	0.016	3.31	0.001	0.022	0.085	***
hhszgr3	0.010	0.014	0.77	0.442	-0.016	0.037	
hhszgr4	0.000	0.013	0.03	0.978	-0.025	0.026	
hhszgr5	0.009	0.014	0.62	0.534	-0.018	0.035	
hhszgr6	0.008	0.015	0.54	0.587	-0.021	0.037	
hhszgr7	-0.014	0.017	-0.81	0.417	-0.048	0.020	
Constant	1.387	0.035	40.16	0.000	1.319	1.454	***

Mean dependent var	0.579	SD dependent var	0.157
R-squared	0.269	Number of observations	2142.000
F-test	52.074	Prob > F	0.000
Akaike crit. (AIC)	-2496.655	Bayesian crit. (BIC)	-2405.943

*** $p<0.01$, ** $p<0.05$, * $p<0.1$

$$\hat{Y} = 1.387 - 0.078^* X_2 - 0.011^*n2 - 0.004^*n3 - 0.117^*n4 - 0.047^*n5 - 0.080^*n7 + 0.013^*n8 + 0.117^*hhszgr1 + 0.054^* hhszgr2 + 0.010^* hhszgr3 + 0.009^* hhszgr5 + 0.008^* hhszgr6 - 0.014^*hhszgr7$$

Note- Here, the hhszgr are the dummy variables.

Interpretation of the coefficients of the dummy variables:

Here, the reference group is the group of households with members greater than or equals 8. The intercept term captures the characteristics of the base group i.e., the average food-share of the base group and it is highly significant.

In case of the other coefficients of dummy variables, it measures the difference between food-share of the corresponding group and the base group, on an average and keeping other factors constant. Among them, the coefficients of s1 and s2 (household size – 1 and 2) are statistically significant i.e., there is a difference in average food share between very small households and very large households. But, as household size exceeds 3, the difference becomes insignificant on an average, given our sample.

The results satisfy the Deaton-Paxon puzzle. For smaller households, the food-share is relatively more and it decreases with increase in household size.

Q 3 b) Now, we have to compare the dummy variable model with the model in specification 2(c). We see that the dependent variable is the same but the independent variables appear slightly differently. So, we first check the adjusted R-squared. But both of them are very close to each other about 0.26 approximately. So, we check the AIC and find that the non-dummy model has lower AIC and thus a better model.

For a better comparison we use the rvf plots and also the sktest for normality.

Here, we find that the rvf plots and the sktest shows similar results in both models.

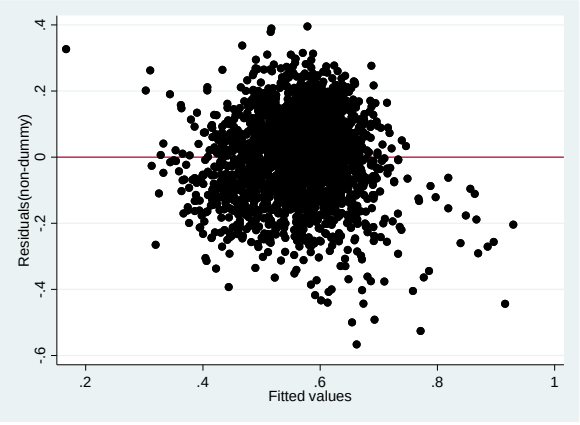

Fig- rvf plot (non-dummy)

Skewness/Kurtosis tests for Normality of non-dummy model

			------ joint ------		
Variable	Obs.	P(Skewness)	P(Kurtosis)	adj_chi2(2)	Prob>chi2
e(non-dummy)	2,142	0.000	0.021	43.720	0.000

Q 4. Here we perform the Chow Test to check if the model has different specifications for rural and urban areas.

Firstly, we have estimated regression function assuming no different specifications for rural and urban area and obtain the restricted RSS_3.

Linear regression – for no different specifications.

food share	Coef.	St. Error	t-value	p-value	(95% Conf	Interval)	Sig
lnapci	-0.078	0.003	-25.33	0.000	-0.084	-0.072	***
lnhhsz	-0.033	0.007	-4.55	0.000	-0.047	-0.019	***
n2	-0.011	0.026	-0.42	0.672	-0.062	0.040	
n3	-0.002	0.024	-0.07	0.944	-0.048	0.045	
n4	-0.101	0.026	-3.85	0.000	-0.152	-0.049	***
n5	-0.040	0.028	-1.42	0.155	-0.095	0.015	
n6	-0.079	0.030	-2.63	0.009	-0.138	-0.020	***
n7	-0.055	0.024	-2.33	0.020	-0.102	-0.009	**
n8	0.030	0.023	1.34	0.179	-0.014	0.074	
Constant	1.420	0.034	41.19	0.000	1.352	1.487	***

Mean dependent var	0.571	SD dependent var	0.156
R-squared	0.247	Number of observations	2418.000
F-test	87.999	Prob > F	0.000
Akaike crit. (AIC)	-2776.764	Bayesian crit. (BIC)	-2718.857

*** $p<0.01$, ** $p<0.05$, * $p<0.1$

The restricted $RSS_1=RSS_R$ = 44.5305457 with df = 2408

$$\hat{Y} = 1.420 - 0.078 * \ln(apci) - 0.033 * lnhhsz - 0.011 * n2 - 0.002 * n3 - 0.101 * n4 - 0.040 * n5 - 0.079 * n6 - 0.055 * n7 + 0.030 * n8$$

Secondly, we estimate the model only for rural and urban areas separately to obtain the Unrestricted residual sum of squares (RSS_{UR}) = $RSS_2 + RSS_3$

Linear regression for rural area

food share	Coef.	St Error	t-value	p-value	[95% Conf	Interval]	Sig
lnapci	-0.027	0.005	-5.45	0.000	-0.036	-0.017	***
lnhhsz	-0.024	0.010	-2.39	0.017	-0.044	-0.004	**
n2	-0.017	0.034	-0.50	0.619	-0.084	0.050	
n3	-0.036	0.033	-1.09	0.276	-0.101	0.029	
n4	-0.114	0.036	-3.13	0.002	-0.185	-0.042	***
n5	-0.078	0.037	-2.09	0.037	-0.151	-0.005	**
n6	-0.106	0.039	-2.68	0.007	-0.183	-0.028	***
n7	-0.018	0.034	-0.52	0.601	-0.084	0.049	
n8	0.065	0.032	2.04	0.042	0.002	0.128	**
Constant	0.973	0.052	18.60	0.000	0.870	1.075	***

Mean dependent var	0.633	SD dependent var	0.141
R-squared	0.044	Number of observations	1274.000
F-test	6.455	Prob > F	0.000
Akaike crit. (AIC)	-1416.670	Bayesian crit. (BIC)	-1365.170

*** *p<0.01, ** p<0.05, * p<0.1*

The RSS for rural model is RSS_2 = 24.1517203 with df = 1264

Linear regression for urban areas

food share	Coef.	St Error	t-value	p-value	[95% Conf	Interval]	Sig
lnapci	-0.096	0.004	-21.89	0.000	-0.105	-0.088	***
lnhhsz	-0.022	0.009	-2.36	0.019	-0.041	-0.004	**
n2	-0.009	0.036	-0.24	0.808	-0.079	0.061	
n3	0.080	0.032	2.52	0.012	0.018	0.142	**
n4	-0.035	0.035	-1.00	0.318	-0.103	0.033	
n5	-0.004	0.039	-0.11	0.910	-0.080	0.072	
n6	-0.065	0.042	-1.54	0.123	-0.148	0.018	
n7	-0.077	0.030	-2.57	0.010	-0.135	-0.018	**
n8	0.041	0.029	1.42	0.155	-0.016	0.098	
Constant	1.510	0.052	28.95	0.000	1.408	1.613	***

Mean dependent var	0.503	SD dependent var	0.144
R-squared	0.322	Number of observations	1144.000
F-test	59.726	Prob > F	0.000
Akaike crit. (AIC)	-1607.228	Bayesian crit. (BIC)	-1556.806

*** *p<0.01, ** p<0.05, * p<0.1*

The RSS for urban areas is RSS_3 = 16.1513985 with d.f = 1134

Since, the two samples are assumed independent by Chow Test, $RSS_{UR} = RSS_2 + RSS_3$ = 40.3031188

The idea behind Chow Test is that if in fact there is no structural breaks, then RSS_{UR} and RSS_R should not be statistically different.

Under null hypothesis of parameter stability,

$F = ((RSS_R - RSS_{UR})/k)/((RSS_{UR})/n1+n2-2k)) \sim F_{k,(n1+n2-2k)}$

Finally, we do not reject the null hypothesis if the computed F- value does not exceed the critical F-value from the F-table at a given level of significance(here, 5% level of significance)

Here, F-calculated = 25.152817 > $F_{10,2408}$ = 1.8346

Hence, reject the null hypothesis of parameter stability in the food share-per capita income, household size, composition between rural and urban areas on the basis of our sample data.