Regression Models with Dummy Independent Variables

- Dummy variables classify data into mutually exclusive categories
- Regression model with only dummy or qualitative variables – Analysis of Variance (ANOVA) – compares the mean of two or more categories
- Regression model with mix of qualitative and quantitative variables – Analysis of Covariance (ANCOVA) – examines the main and interaction effects of categorical variables on a continuous dependent variable, controlling the effects of other continuous variables
- In addition to examining impact of qualitative aspects or attributes, dummy (independent) variables are also used for seasonality analysis and examining structural breaks/differences

ANOVA: Example 1

Model Specification	Interpretation
To examine if average monthly per capita consumption expenditure (MPCE) varies depending on whether the households belong to rural, urban and semi-urban areas, i.e., $Y_i = \alpha + \beta_1 D_{1i} + \beta_2 D_{2i} + u_i$ Y_i = Monthly per capita consumption expenditure Household from three different types of locations D_{1i} = 1 if the household is in urban area D_{1i} = 0 otherwise D_{2i} = 1 if the household is in semi-urban area D_{2i} = 0 otherwise Base category = Households from **rural area**	**Three Alternatives:** (1) If D_{1i} = 0, D_{2i} = 0 => $E(Y_i) = \alpha$ (2) If D_{1i} = 1, D_{2i} = 0 => $E(Y_i) = \alpha + \beta_1$ (3) If D_{1i} = 0, D_{2i} = 1 => $E(Y_i) = \alpha + \beta_2$ **Possibilities** (a) ***Comparison between rural and urban households*** (1) B_1 is not statistically significant => average MPCE of urban households is not significantly different from that of rural households (2) B_1 is statistically significant and positive => average MPCE of urban households is significantly higher than that of rural households (3) B_1 is statistically significant and negative => average MPCE of urban households is significantly lower than that of rural households (b) ***Comparison between rural and semi-urban households*** (1) B_2 is not statistically significant => average MPCE of semi-urban households is not significantly different from that of rural households (2) B_2 is statistically significant and positive => average MPCE of semi-urban households is significantly higher than that of rural households (3) B_2 is statistically significant and negative => average MPCE of semi-urban households is significantly lower than that of rural households ***(c) Comparison between urban and semi-urban households*** (1) B_1 is not significantly different from B_2 => average MPCE of urban households is not significantly different from that of semi-urban households (2) B_1 is significantly higher that B_2 => average MPCE of urban households is significantly higher than that of semi-urban households (3) B_1 is significantly lower than B_2 => average MPCE of urban households is significantly lower than that of semi-urban households

ANOVA: Example 2

Model Specification	Interpretation
To examine if monthly per capita consumption expenditure varies depending on (i) if the households belong	**Four Alternatives:** (1) If D_{1i} = 0, D_{2i} = 0 => $E(Y_i) = \alpha$ (Rural, Female Head)

to rural or urban areas, and (ii) if the household head is male or female, i.e., $Y_i = \alpha + \beta_1 D_{1i} + \beta_2 D_{2i} + \beta_3 (D_{1i} * D_{2i}) + u_i$ Y_i = Monthly per capita consumption expenditure D_{1i} = 1 if the household is in urban area D_{1i} = 0 otherwise D_{2i} = 1 if the household head is male D_{2i} = 0 otherwise Base category = Households from **<u>rural area</u>** with **<u>female head</u>**	(2) If D_{1i} = 1, D_{2i} = 0 => $E(Y_i) = \alpha + \beta_1$ (Urban, Female Head) (3) If D_{1i} = 0, D_{2i} = 1 => $E(Y_i) = \alpha + \beta_2$ (Rural, Male Head) (4) If D_{1i} = 1, D_{2i} = 1 => $E(Y_i) = \alpha + \beta_1 + \beta_2 + \beta_3$ (Urban, Male Head) **<u>Some Possibilities</u>** ***(a) Differential impact of being in urban area (with female head)*** (1) B_1 is not statistically significant => average MPCE of urban households female head is not significantly different from that of rural households (2) B_1 is statistically significant and positive => average MPCE of urban households with female head is significantly higher than that of others (3) B_1 is statistically significant and positive => average MPCE of urban households with female head is significantly lower than that of others ***(b) Differential impact of having male household head (in rural areas)*** (1) B_2 is not significant => average MPCE of rural households with male head is not significantly different from those with female head (2) B_2 is significant and positive => average MPCE of rural households with male head is significantly higher than those with female head (3) B_2 is significant and negative => average MPCE of rural households with male head significantly lower than those with female head ***(c) Differential impact of having male head in urban area*** (1) B_3 is not statistically significant => average MPCE of urban households with male head is not significantly different from others (2) B_3 is statistically significant and positive => average MPCE of urban households with male head is significantly higher than that of others (3) B_3 is statistically significant and negative => average MPCE of urban households with male head is significantly lower than that of others

<u>ANOVA: Example 3</u>

Model Specification	Interpretation
To examine if monthly MPCE varies depending on (i) if the households belong to rural or urban areas, (ii) if the household head is male or female, and (iii) if the household is of APL or BPL category, i.e., $Y_i = \alpha + \beta_1 D_{1i} + \beta_2 D_{2i} + \beta_3 D_{3i} + \beta_4 (D_{1i} * D_{2i}) +$ $\beta_5 (D_{1i} * D_{3i}) + \beta_6 (D_{2i} * D_{3i}) + \beta_7 (D_{1i} * D_{2i} * D_{3i}) + u_i$ Y_i = Monthly per capita consumption expenditure D_{1i} = 1 if the household is in urban area D_{1i} = 0 otherwise D_{2i} = 1 if the household head is male D_{2i} = 0 otherwise D_{3i} = 1 if the household is APL D_{3i} = 0 otherwise Base category = **<u>BPL</u>** households from **<u>rural area</u>** with **<u>female head</u>**	**Eight Alternatives:** (1) If D_{1i} = 0, D_{2i} = 0, D_{3i}=0 => $E(Y_i) = \alpha$ (Rural, Female Head, BPL) (2) If D_{1i} = 1, D_{2i} = 0, D_3i=0 => $E(Y_i) = \alpha + \beta_1$ (Urban, Female Head, BPL) (3) If D_{1i} = 0, D_{2i} = 1, D_{3i}=0 =>, $E(Y_i) = \alpha + \beta_2$ (Rural, Male Head, BPL) (4) If D_{1i} = 0, D_{2i} = 0, D_{3i}=1 =>, $E(Y_i) = \alpha + \beta_3$ (Rural, Female Head, APL) (5) If D_{1i} = 1, D_{2i} = 1, D_{3i}=0 => $E(Y_i) = \alpha + \beta_1 + \beta_2 + \beta_4$ (Urban, Male Head, BPL) (6) If D_{1i} = 1, D_{2i} = 0, D_{3i}=1 => $E(Y_i) = \alpha + \beta_1 + \beta_3 + \beta_5$ (Urban, Female Head, APL) (7) If D_{1i} = 0, D_{2i} = 1, D_{3i}=1 => $E(Y_i) = \alpha + \beta_2 + \beta_3 + \beta_6$ (Rural, Male Head, APL) (8) If D_{1i} = 1, D_{2i} = 1, D_{3i}=1 => $E(Y_i) = \alpha + \beta_1 + \beta_2 + \beta_3 + \beta_4 + \beta_5 + \beta_6 + \beta_7$ (Urban, Male Head, APL)

ANCOVA: Example 1

Model Specification	Interpretation
To examine if monthly per capita consumption expenditure varies depending on income and whether the households belong to rural or urban areas, i.e., $Y_i = \alpha + \beta_1 D_{1i} + \beta_2 X_i + \beta_3 (D_{1i} * X_i) + u_i$ Y_i = Monthly per capita consumption expenditure X_i = Monthly income of the household Household from two different types of location – rural and urban D_{1i} = 1 if the household is in urban area D_{1i} = 0 otherwise Base category = Households from **rural area**	Two Alternatives: (1) Given X_i and D_{1i} = 0, $E(Y_i) = \alpha + \beta_2 X_i$ (2) Given X_i and D_{1i} = 1, $E(Y_i) = (\alpha + \beta_1) + (\beta_2 + \beta_3) X_i$ The PRFs will differ depending on statistical significance and sign of β_1 and β_3 **Possibilities** (1) If both β_1 and β_3 are not significant, the two PRFs will coincide (2) If β_1 is significant and β_3 not, the two PRFs will be parallel (difference will be only in respect of intercept – autonomous consumption) (a) If β_1 is positive, PRF for urban households will have higher intercept (b) If β_1 is negative, PRF for urban households will have lower intercept (3) If β_3 is significant and β_1 not, the two PRFs will be concurrent (difference will be only in respect of slope – induced consumption) (a) If β_3 is positive, PRF for urban households will be steeper (b) If β_3 is negative, PRF for urban households will be flatter (4) If both β_1 and β_3 are significant, the two PRFs will be dissimilar (a) If both β_1 and β_3 positive, PRF for urban households will be steeper with a higher intercept (b) If both β_1 and β_3 negative, PRF for urban households will be flatter with a lower intercept (c) If β_1 is positive but and β_3 is negative, PRF for urban households will be flatter with a higher intercept (d) If β_1 is negative but and β_3 is positive, PRF for urban households will be steeper with a lower intercept

ANCOVA: Example 2

Model Specification	Interpretation
To examine if MPCE varies depending on income, the households belong to rural or urban areas, and if the household head is male or female, i.e., $Y_i = \alpha + \beta_1 D_{1i} + \beta_2 D_{2i} + \beta_3 (D_{1i} * D_{2i}) +$ $\beta_4 X_i + \beta_5 (D_{1i} * X_i) + \beta_6 (D_{2i} * X_i) + u_i$ Y_i = Monthly per capita consumption expenditure X_i = Monthly income of the household Household from two different types of location – rural and urban D_{1i} = 1 if the household is in urban area D_{1i} = 0 otherwise D_{2i} = 1 if the household head is male D_{2i} = 0 otherwise Base category = Households from **rural area** with **female head**	Four Alternatives: (1) Given X_i and D_{1i} = 0, D_{2i}=0: $E(Y_i) = \alpha + \beta_4 X_i$ (2) Given X_i and D_{1i} = 1, D_{2i}=0: $E(Y_i) = (\alpha + \beta_1) + (\beta_4 + \beta_5) X_i$ (3) Given X_i and D_{1i} = 0, D_{2i}=1: $E(Y_i) = (\alpha + \beta_2) + (\beta_4 + \beta_6) X_i$ (4) Given X_i and D_{1i} = 1, D_{2i}=1: $E(Y_i) = (\alpha + \beta_1 + \beta_2 + \beta_3) + (\beta_4 + \beta_5 + \beta_6) X_i$ The PRFs will differ depending on statistical significance and sign of β_1, β_2, β_3, β_4, β_5, and β_6 **Possibilities** **How will you explain the coefficients of alternative PRFs?**

Two Alternative Forms: (i) $Y_i = \alpha + \beta_1 D_{1i} + \beta_2 D_{2i} + u_i$, and (ii) $Y_i = \gamma_1 D_{1i} + \gamma_2 D_{2i} + \gamma_3 D_{3i} + v_i$

How will the results and the interpretation of the coefficients differ?

Empirical Example:

- Cross-sectional data set of 100 households
- **Dependent Variable:** (i) Monthly per capita consumption expenditure (LNMPCE)*
  Independent Variables: (ii) Age of the household head (LNHHAGE)*, (iii) Per capita landholding (PCLAND), (iii) Household size (LNHHSIZE)*, (iv) Location of the households (RURAL, URBAN and SEMIURBAN), (vi) If the household is below the poverty line (BPL)
- **Two interaction terms**: (i) Rural households below the poverty line (RURBPL=RURAL*BPL), and (ii) Per capita landholding of rural households (RURLAND=RURAL*PCLAND)
- Estimations of alternative models with URBAN as the base (reference) category
- Shortlisting of the following two models initially based on R^2:

*These three variables are measured in natural logarithmic scale

Model I

Number of Observation				100
F(8, 91)				59.11
Prob. > F-Stat				< 0.001
R-squared				0.363
Root MSE				0.50841
Variable	Coefficient	Robust Std. Err.	t-Stat	P>t
SEMIURBAN	0.17379	0.14038	1.24	0.219
RURAL	0.24295	0.15306	1.59	0.116
LNHHSIZE	-0.61185	0.10184	-6.01	<0.001
LNHHAGE	0.22987	0.19642	1.17	0.245
BPL	-0.18597	0.13794	-1.35	0.181
RURBPL	0.09078	0.19716	0.46	0.646
RURLAND	-0.42537	0.84882	-0.5	0.617
PCLAND	2.17516	0.75790	2.87	0.005
Intercept	7.64819	0.72726	10.52	<0.001

Model II

Number of Observation				100
F(6, 93)				70.53
Prob. > F-Stat				<0.001
R-squared				0.3519
Root MSE				0.50728
Variable	Coefficient	Robust Std. Err.	t-Stat	P>t
SEMIURBAN	0.1379	0.1394	0.99	0.325
RURAL	0.2690	0.1110	2.42	0.017
LNHHSIZE	-0.6099	0.0970	-6.28	0.001
BPL	-0.1585	0.1098	-1.44	0.152
RURLAND	-0.4974	0.6963	-0.71	0.477
PCLAND	2.2442	0.6875	3.26	0.002
Intercept	8.5250	0.1487	57.32	0.001

Comparison between the two models:

- Not much difference in R^2 between the two models (to be tested statistically)
- More variables turnout to be statistically significant in Model II
- No change in sign of the statistically significant coefficients
- **Finally, selection of Model II for further discussions**

Interpretation of the Results

- Coefficient of RURAL is significant and positive => Average MPCE of rural households is higher than that of urban households.
- Coefficient of SEMIURBAN is not significant => Average MPCE of semi-urban households is not significantly different from that of the urban households.
- Coefficient of LNHHSIZE is statistically significant and negative => Households with more members in the family have lower average MPCE.
- Coefficient of PCLAND is statistically significant and positive => Households with more landholding per member in the family have higher average MPCE.
- Coefficient of BPL is not statistically significant=>Average MPCE does not differ depending on whether households below to the BPL category

Use of Dummy Variables for Seasonality Analysis

- Assumption: The components of the time-series is additive, i.e.,

  TS = Trend (T) + Seasonal (S) + Cyclical (C) + Randomness (U)

  ✓ **<u>Steps to be followed:</u>**

  - Use of dummy for every quarter (without intercept) or for three quarters (with intercept) treating the omitted quarter as the base or reference
  - Estimation of the residuals – deseasonalized values of the time-series

- **<u>Important Questions</u>**:

  ✓ Will the results differ depending on selection of the base/reference quarter?
  <u>Answer:</u> NO – The deseasonalized time-series will be the same irrespective of selection of the base/reference quarter

  ✓ Consider the function: GDP = f(GFCF). How to account for seasonality in GFCF, if any?

  <u>Frisch-Waugh Theorem:</u> Use of dummy variables in GDP = f(GFCF) will deseasonalize both GDP and GFCF

  - Running regression of GDP on the dummy variables and estimating the residuals (U1)
  - Running regression of GFCF on the dummy variables and estimating the residuals (U2)
  - Regressing U1 on U2 – Same coefficient of GFCF as one gets when GDP is regressed on GFCF and the dummy variables

  ✓ What to do when the components of a time-series are multiplicative? – To be discussed in time-series econometrics

Seasonality Analysis: Example									
SUMMARY OUTPUT (Model I)					SUMMARY OUTPUT (Model II)				
Multiple R	0.261				Multiple R	0.261			
R Square	0.068				R Square	0.068			
Adjusted R Square	-0.049				Adjusted R Square	-0.049			
Standard Error	0.155				Standard Error	0.155			
Observations	28				Observations	28			
	df	*SS*	*MS*	*F*		*df*	*SS*	*MS*	*F*
Regression	3	0.042161	0.014054	0.582704	Regression	3	0.042161	0.014054	0.582704
Residual	24	0.578828	0.024118		Residual	24	0.578828	0.024118	
Total	27	0.620988			Total	27	0.620988		
	Coefficients	*Standard Error*	*t Stat*	*P-value*		*Coefficients*	*Standard Error*	*t Stat*	*P-value*
Intercept	14.923	0.058698	254.231	1.1E-42	Intercept	14.824	0.058698	252.5	1.29E-42
D1	-0.0989	0.083011	-1.19149	0.24511	D2	0.0085	0.083011	0.1	0.919106
D2	-0.0904	0.083011	-1.08886	0.287029	D3	0.0402	0.083011	0.5	0.63239
D3	-0.0587	0.083011	-0.70695	0.486408	D4	0.0989	0.083011	1.2	0.24511

Model I						**Model II**					
Obs.	Predicted GDP	Residuals	Obs.	Predicted GDP	Residuals	Obs.	Predicted GDP	Residuals	Obs.	Predicted GDP	Residuals
1	14.824	-0.21485	15	14.864	-0.00601	1	14.824	-0.21485	15	14.864	-0.00601
2	14.832	-0.22893	16	14.923	0.003556	2	14.832	-0.22893	16	14.923	0.003556
3	14.864	-0.19649	17	14.824	0.084883	3	14.864	-0.19649	17	14.824	0.084883
4	14.923	-0.20196	18	14.832	0.086258	4	14.923	-0.20196	18	14.832	0.086258
5	14.824	-0.15017	19	14.864	0.066738	5	14.824	-0.15017	19	14.864	0.066738
6	14.832	-0.15386	20	14.923	0.07161	6	14.832	-0.15386	20	14.923	0.07161
7	14.864	-0.13404	21	14.824	0.143054	7	14.864	-0.13404	21	14.824	0.143054
8	14.923	-0.146	22	14.832	0.151753	8	14.923	-0.146	22	14.832	0.151753
9	14.824	-0.07783	23	14.864	0.140815	9	14.824	-0.07783	23	14.864	0.140815
10	14.832	-0.07596	24	14.923	0.149779	10	14.832	-0.07596	24	14.923	0.149779
11	14.864	-0.07551	25	14.824	0.219563	11	14.864	-0.07551	25	14.824	0.219563
12	14.923	-0.08343	26	14.832	0.219427	12	14.923	-0.08343	26	14.832	0.219427
13	14.824	-0.00465	27	14.864	0.204511	13	14.824	-0.00465	27	14.864	0.204511
14	14.832	0.001316	28	14.923	0.206451	14	14.832	0.001316	28	14.923	0.206451

Testing for Structural Break

(1) Two sub-periods, i.e., one possible structural break
(2) Regression equation for the first sub-period: $Y_i = \alpha_1 + \beta_1 X_i + \gamma_1 Z_i + u_{1i}$
(3) Regression equation for the second sub-period: $Y_i = \alpha_2 + \beta_2 X_i + \gamma_2 Z_i + u_{2i}$
(4) Combined model: $Y_i = \alpha_1 + (\alpha_2 - \alpha_1)D_{1i} + \beta_1 X_i + (\beta_2 - \beta_1)D_{2i} + \gamma_1 Z_i + (\gamma_2 - \gamma_1)D_{3i} + u_i$
(Unrestricted model)

(i) D_1 = 1 for the second sub-period, and D_1=0 for the first sub-period
(ii) D_2 = X for observations corresponding to the second sub-period, and D_2=0 for observations corresponding to the first sub-period
(iii) D_3 = Z for observations corresponding to the second sub-period, and D_3=0 for observations corresponding to the first sub-period

Restrictions:

	Elimination of Dummy
All coefficients are the same $\alpha_1 = \alpha_2; \beta_1 = \beta_2; \gamma_1 = \gamma_2$	D_1, D_2, D_3
Only intercepts change: $\beta_1 = \beta_2; \gamma_1 = \gamma_2$	D_2, D_3
Only coefficients of X change: $\alpha_1 = \alpha_2; \gamma_1 = \gamma_2$	D_1, D_3
Only coefficients of Z change: $\alpha_1 = \alpha_2; \beta_1 = \beta_2$	D_1, D_2
Only slopes change: $\alpha_1 = \alpha_2$	D_1
Only intercepts and coefficients of X change: $\gamma_1 = \gamma_2$	D_3
Only intercepts and coefficients of Z change: $\beta_1 = \beta_2$	D_2
All the slope coefficients and the intercept change	None (Full Model)