Dummy Variables

- Dummy variables are recoded nominal or ordinal variables
  - Coded into dichotomous variables
  - If original variables has *k* attributes, you create (*k* – 1) dummy variables

Why?

- Consider "ethnicity"—if coded 1=White, 2=AA, 3=Latino, etc., then regression sees this as a continuous variable, which is not accurate. It's a categorical (nominal) variable.
- Why *k*-1? Because we don't need to create dummy variables for all the original attributes. The analysis treats the missing dummy variable as a baseline with which to compare all others. (If you did code all attributes and tried to run the multivariate analysis, your analysis would be in error.)

How it's done

- Consider the variable "ethnicity" with five attributes:
  1. White,
  2. African-American,
  3. Latino,
  4. Asian/Pacific Islander,
  5. and Other

“Ethnicity” Before Recoding

Subject.ID	*Ethnicity* 1=White 2=Latino 3=Afr Amer 4=Asian/PI 5=Other
1	3
2	3
3	1
4	4
5	1
6	2
7	2
8	5
9	2
10	2

Recode into four dichotomous variables:

1. “White” 1 = White; 0 = Not White
2. “African-American” 1 = AA; 0 = not AA
3. Latino 1 = Latino; 0 = Not Latino
4. Asian/PI 1 = Asian/PI; 0 = Not Asian/PI

After Recoding

Subject.ID	*White* 0=non-White 1=White	*Latino* 0=non-Latino 1=Latino	*Afr.amer* 0=non-AA 1=AA	*Asian.PI* 0=non-Asian.PI 1=Asian.PI
1	0	0	1	0
2	0	0	1	0
3	1	0	0	0
4	0	0	0	1
5	1	0	0	0
6	0	1	0	0
7	0	1	0	0
8	0	0	0	0
9	0	1	0	0
10	0	1	0	0

"Other" is the baseline.
But is it really missing?
Look at subject #8