# METHODOLOGICAL ISSUES AND STRATEGIES FOR USING AND MEASURING RACE IN REGRESSION ANALYSES

**Melody S. Goodman, PhD**
**Associate Professor of Biostatistics**
**Associate Dean for Research**
**NYU School of Global Public Health**

NYU SCHOOL OF GLOBAL PUBLIC HEALTH

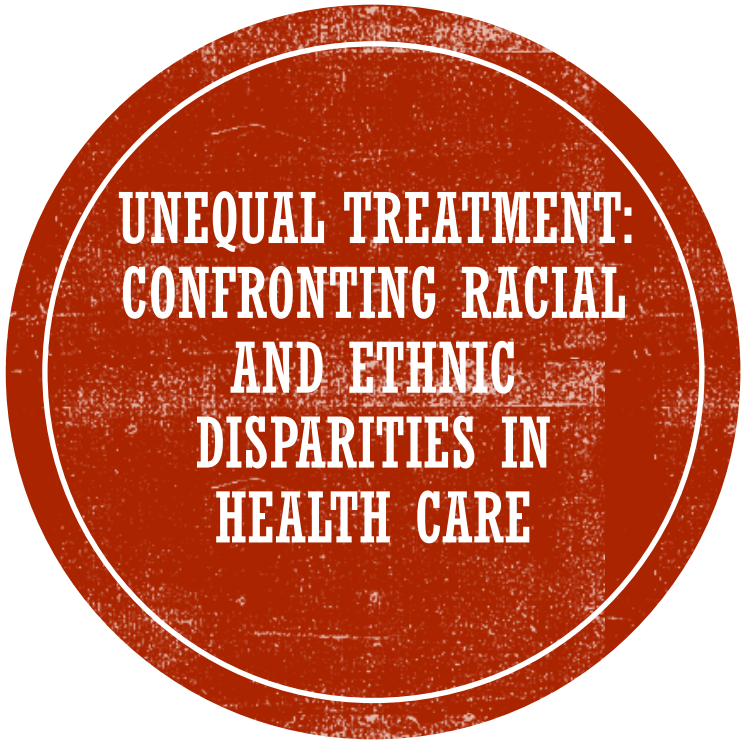

# INTRODUCTION - CONTEXT

# HEALTH DISPARITIES

- Differences in health that:
  - Are systematic and plausibly avoidable
  - Influenced by policies
  - Put socially disadvantaged groups at further disadvantage with respect to their health

- NIH definition of health disparities:
  - "Differences in the incidence, prevalence, mortality, and burden of diseases and other adverse health conditions that exist among specific population groups in the U.S."

# HEALTHCARE DISPARITIES




"Racial or ethnic differences in the quality of healthcare that are <u>not</u> due to access-related factors or clinical needs, patient preferences, and appropriateness of intervention."

Institute of Medicine Committee on Understanding and Eliminating Disparities in Health Care

# RACE ON BIRTH RECORDS BEFORE 1989

| Father | Mother | Baby* |
|---|---|---|
| White | White | White |
| Black | Black | Black |
| Hawaiian | Hawaiian | Hawaiian |
| Asian | Asian | Asian |
| White | Black | Black |
| Black | White | Black |
| Black | Asian | Black |
| White | Asian | Asian |
| Black | Hawaiian | Hawaiian |

*Not reported on birth records but used for other statistical purposes

**Source**: http://www.cdc.gov/nchs/data/techap99.pdf

# RACE: WHAT IT IS, WHAT IT IS NOT

"Racial categories reflect social and ideological conventions, not meaningful natural distinctions"

**Race is a man-made social construct**

**Source**: Krieger N. Discrimination and Health Inequities. In: Berkamn LF, Kawachi I, Glymour MM, editors. *Soc. Epidemiol.* Second Edition. New York: Oxford University Press; 2014.

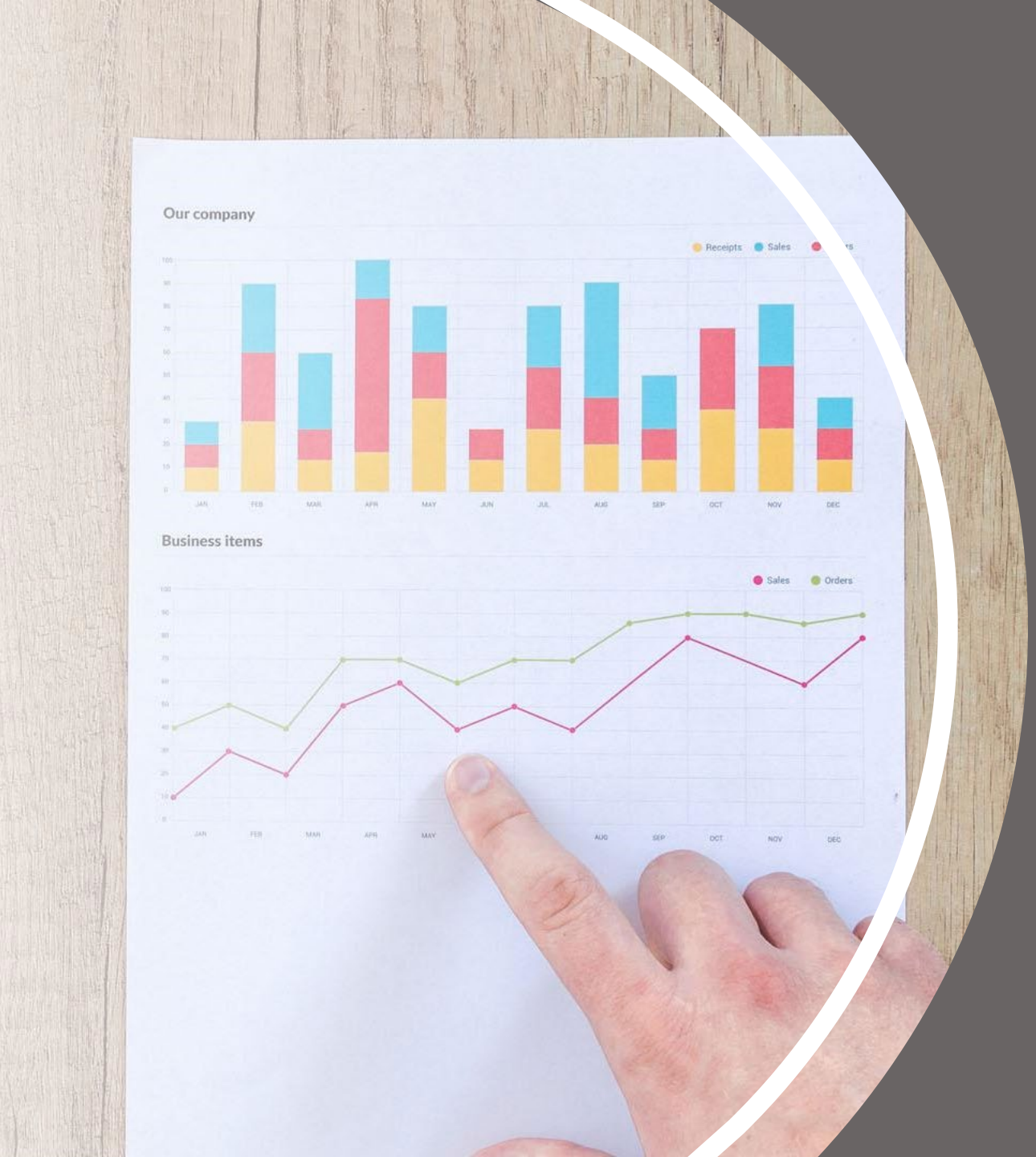
# SALIENCE OF RACE IN STATISTICAL ANALYSIS

Race is salient to statistical analysis because:

- Early science → flawed racial thinking (e.g., race as biology).
  - This thinking continues to creep into contemporary discourses and research

- Vulnerable populations include disproportionate numbers of racial and ethnic minorities
  - Minorities have poorer health and worse access to care
  - The interventions that address these problems often are colorblind—that is, they provide individuals with immediate care, but they are not intended to address broader racial factors that may undergird the disparities.

- Racialization is salient whenever the purpose of research is to understand how racial factors influence disease distributions
  - Research on racial differences may be either problematic or constructive
  - To advance racist hypotheses; many foundational statistical methods were generated by eugenicists
  - To contest racist research
  - To identify social determinants of disparities




# DATA DRIVEN APPROACHES

*"Data never speak for themselves. It is the questions we pose (and those we fail to ask) as well as our theories, concepts and ideas that bring a narrative and meaning to marginal distributions, correlations, regression coefficients, and statistics of all kinds."*

*- Lawrence Bobo 2004*

9

# CODING RACE/ETHNICITY VARIABLES IN REGRESSION ANALYSIS

# HOW IS RACE MEASURED?

- Self-report
- Other-report
  - Visual
  - Objective
- Familial factors
- Nativity
- Identity
- Social class

**Sample Survey Question**

Which of the following best describes your race?

1. African American or Black
2. Asian American or Asian
3. White or Caucasian
4. Other

From: NYC Low-Income Housing, Neighborhoods and Health Study

# RACE IS A CATEGORICAL (NOMINAL) VARIABLE

- Categorical variables require special attention in regression analysis because, unlike dichotomous or continuous variables, they cannot by entered into the regression equation just as they are
- Categorical variables need to be recoded into a series of variables which can be entered into the regression model
- There are a variety of coding systems that can be used when coding categorical variables
- Ideally, you would choose a coding system that reflects the comparisons that you want to make
- We often recode categorical variables in regression analysis using dummy variables
  - Not the only coding scheme that you can use

**Source**: https://stats.idre.ucla.edu/stata/webbooks/reg/chapter5/regression-with-statachapter-5-additional-coding-systems-for-categorical-variables-in-regressionanalysis/

# CHOSE YOUR CODING SYSTEM APPROPRIATELY

- By deliberately choosing a coding system, you can obtain comparisons that are most meaningful for testing your hypotheses.
- If you want to compare each level to the next higher level, you will want to use "forward difference" coding
- Regardless of the coding system you choose, the test of the overall effect of the categorical variable (i.e., the overall effect of **race**) will remain the same.

**Source**: https://stats.idre.ucla.edu/stata/webbooks/reg/chapter5/regression-with-statachapter-5-additional-coding-systems-for-categorical-variables-in-regressionanalysis/

# TYPES OF CONTRASTS & THE COMPARISON MADE

| Name of contrast | Comparison made |
|---|---|
| Simple Coding | Compares each level of a variable to the reference level |
| Forward Difference Coding | Adjacent levels of a variable (each level minus the next level) |
| Backward Difference Coding | Adjacent levels of a variable (each level minus the prior level) |
| Deviation Coding | Compares deviations from the grand mean |
| Statistical Analyst Defined Coding | Statistical Analyst defined contrast |

**Choose a coding system that yields comparisons that make the most sense for testing your hypotheses**



**Source**: https://stats.idre.ucla.edu/stata/webbooks/reg/chapter5/regression-with-statachapter-5-additional-coding-systems-for-categorical-variables-in-regressionanalysis/

# NOTES ON CODING SYSTEMS

- Represent planned comparisons and not *post hoc* comparisons
  - They are comparisons that you plan to do before you begin analyzing your data
  - Not comparisons that you think of once you have seen the results of preliminary analyses

- Some forms of coding make more sense with ordinal categorical variables than with nominal categorical variables.

- Because simple effect coding compares the mean of the dependent variable for each level of the categorical variable to the mean of the dependent variable at for the reference level, it makes sense with a nominal variable.

- However, it may not make as much sense to use a coding scheme that tests the linear effect of **race**.



**Source**: https://stats.idre.ucla.edu/stata/webbooks/reg/chapter5/regression-with-statachapter-5-additional-coding-systems-for-categorical-variables-in-regressionanalysis/

# EXAMPLE DATASET

- Data from NYC community health survey
- https://www1.nyc.gov/site/doh/data/data-sets/community-health-survey-public-use-data.page

- We will focus on the categorical variable **newrace**, which has four levels (1 = Non-Hispanic White, 2 = Black (non-Hispanic), 3 = Hispanic and 4 = Asian)
  - This is how race was categorized in the dataset (limitation of secondary analysis)
  - we will use **body mass index (bmi)** as our dependent variable

- Although our example uses a variable with four levels, these coding systems work with variables that have more or fewer categories

- No matter which coding system you select, you will always have one fewer recoded variables than levels of the original variable
  - In our example, the race variable has four levels so we will have three new variables
  - A variable corresponding to the final level of the categorical variables would be redundant and therefore unnecessary.



**Source**: https://stats.idre.ucla.edu/stata/webbooks/reg/chapter5/regression-with-statachapter-5-additional-coding-systems-for-categorical-variables-in-regressionanalysis/

```
 Observations:         8,781
    Variables:           147
```

| Variable name | Storage type | Display format | Value label | Variable label |
|---|---|---|---|---|
| cid | double | %12.0g | | Unique identifier across all years of CHS |
| strata | double | %12.0g | | Variance stratum - Collapsed |
| survey | double | %12.0g | | Survey number |
| wt21_dual | double | %12.0g | | CHS 2020 Survey weight |
| wt21_dual_q1 | double | %12.0g | | CHS 2020 Long Survey Weight |
| strata_q1 | double | %12.0g | | Variance stratum - Collapsed half sample |
| qxvers | double | %12.0g | | Questionnaire Version. QXVERS = 1 (LONG VERSION) QXVERS = 2 (SHORT VERSION) |
| mood1 | double | %12.0g | mood1 | Q5.1 During the past 30 days, how often did you feel...So sad that nothing coul |

# EXAMINE THE DATA

Before considering any analyses, let's examine the dataset

**Unweighted:** `tabulate race, summarize(bmi)`

| Race/ethnicity variable of person interviewed | Summary of Body Mass Index (kg / sq in) | | |
|---|---|---|---|
| | Mean | Std. dev. | Freq. |
| White | 26.35302 | 5.62598 | 2,752 |
| Black | 28.834652 | 6.5778025 | 1,759 |
| Hispanic | 28.631466 | 7.052452 | 2,357 |
| Asian/PI | 24.535502 | 4.9037819 | 1,275 |
| Total | 27.264005 | 6.3805962 | 8,143 |

# EXAMINE THE KEY VARIABLES

- Calculate the mean of the dependent variable, **bmi**, for each level of **race**
- This will help in interpreting the output from regression analyses

# EXAMINE THE KEY VARIABLES

```
Survey: Mean estimation

Number of strata =   159                  Number of obs    =     8,143
Number of PSUs   = 8,143                  Population size = 5,988,401
                                          Design df       =     7,984
```

| | Mean | Linearized std. err. | [95% conf. | interval] |
|---|---|---|---|---|
| c.bmi@newrace | | | | |
| White | 26.32592 | .1708285 | 25.99105 | 26.66079 |
| Black | 28.4591 | .2487949 | 27.9714 | 28.94681 |
| Hispanic | 28.43682 | .1997335 | 28.04529 | 28.82835 |
| Asian/PI | 24.32051 | .2310858 | 23.86752 | 24.7735 |

- Calculate the mean of the dependent variable, **bmi**, for each level of **race**
- This will help in interpreting the output from regression analyses

```
svyset cid [pweight=wt21_dual], strata(strata) vce(linearized) singleunit(missing)
```

**Weighted:** `svy: mean bmi, over(newrace)`

# DUMMY CODING

- It is a way to make the categorical variable into a series of dichotomous variables
  - Variables can have a value of zero or one only
  - For all but one of the levels of the categorical variable, a new variable will be created that has a value of one for each observation at that level and zero for all others.

- In our example using the variable race
  - The first dummy variable (x1) will have a value of one for each observation in which race is Black, and zero for all other observations.
  - Likewise, we create **x2** to be 1 when the person is Hispanic, and 0 otherwise
  - **x3** is 1 when the person is Asian, and 0 otherwise.

- The level of the categorical variable that is coded as zero in all of the new variables is the reference level, or the level to which all of the other levels are compared.
  - In our example, white is the reference level because this is what is typically done (doesn't mean it's ideal)
  - You can select any level of the categorical variable as the reference level.

**Source**: https://stats.idre.ucla.edu/spss/faq/coding-systems-for-categorical-variables-in-regression-analysis/

# DUMMY CODING EXAMPLE

| Level of race | New variable 1 (x1) | New variable 2 (x2) | New variable 3 (x3) |
|---|---|---|---|
| 1 (White) | 0 | 0 | 0 |
| 2 (Black) | 1 | 0 | 0 |
| 3 (Hispanic) | 0 | 1 | 0 |
| 4 (Asian) | 0 | 0 | 1 |

- The coefficient for x1 is the mean of the dependent variable for group 1 minus the mean of the dependent variable for the omitted group.
- The coefficient for x1 is the mean of BMI for the Black group minus the mean of BMI for the White group
- The coefficient for x2 would be the mean of BMI for the Hispanic group minus the mean of BMI for the White group
- The coefficient for x3 would be the mean of BMI for the Asian group minus the mean of BMI for the White group.



**Source**: https://stats.idre.ucla.edu/spss/faq/coding-systems-for-categorical-variables-in-regression-analysis/

# DUMMY VARIABLE REGRESSION OUTPUT (UNWEIGHTED)

```
regress bmi i.newrace
```

| Source | SS | df | MS |
|---|---|---|---|
| Model | 20522.7022 | 3 | 6840.90072 |
| Residual | 310954.466 | 8,139 | 38.2054879 |
| Total | 331477.168 | 8,142 | 40.7120079 |

| | | | |
|---|---|---|---|
| Number of obs | = | 8,143 |
| F(3, 8139) | = | 179.06 |
| Prob > F | = | 0.0000 |
| R-squared | = | 0.0619 |
| Adj R-squared | = | 0.0616 |
| Root MSE | = | 6.1811 |

| bmi | Coefficient | Std. err. | t | P>\|t\| | [95% conf. | interval] |
|---|---|---|---|---|---|---|
| newrace | | | | | | |
| Black | 2.481632 | .1886871 | 13.15 | 0.000 | 2.111757 | 2.851507 |
| Hispanic | 2.278446 | .173471 | 13.13 | 0.000 | 1.938398 | 2.618493 |
| Asian/PI | -1.817518 | .2093989 | -8.68 | 0.000 | -2.227994 | -1.407043 |
| _cons | 26.35302 | .1178253 | 223.66 | 0.000 | 26.12205 | 26.58399 |

| Race/ethnicity variable of person interviewed | Summary of B<br>Mean |
|---|---|
| White | 26.35302 |
| Black | 28.834652 |
| Hispanic | 28.631466 |
| Asian/PI | 24.535502 |
| Total | 27.264005 |

# DUMMY VARIABLE REGRESSION OUTPUT (WEIGHTED)

```
svy: regress bmi i.newrace
```

Survey: Linear regression

Number of strata = 159
Number of PSUs = 8,143

Number of obs = 8,143
Population size = 5,988,401
Design df = 7,984
F(3, 7982) = 77.90
Prob > F = 0.0000
R-squared = 0.0533

| bmi | Coefficient | Linearized std. err. | t | P>\|t\| | [95% conf. | interval] |
|---|---|---|---|---|---|---|
| newrace | | | | | | |
| Black | 2.133186 | .3019247 | 7.07 | 0.000 | 1.541335 | 2.725037 |
| Hispanic | 2.110904 | .2628895 | 8.03 | 0.000 | 1.595571 | 2.626236 |
| Asian/PI | -2.00541 | .2876277 | -6.97 | 0.000 | -2.569236 | -1.441585 |
| _cons | 26.32592 | .1708285 | 154.11 | 0.000 | 25.99105 | 26.66079 |

Survey: Mean estimation

Number of strata = 159
Number of PSUs = 8,143

| | Mean |
|---|---|
| c.bmi@newrace | |
| White | 26.32592 |
| Black | 28.4591 |
| Hispanic | 28.43682 |
| Asian/PI | 24.32051 |

# EFFECT CODING

- Other coding systems use more values than just zero and one
- Allows you to make other types of comparisons
- Unlike dummy coding, effect coding allows you to assign different weights to the various levels of the categorical variable.
- The "rule" in dummy coding is that only values of zero and one are valid
- The "rule" in effect coding is that all of the values in any new variable must sum to zero
- Which level is assigned a positive or negative value is not very important
  - 0 1 -1 0 is the same as 0 -1 1 0
  - Both of these codings compare the second and the third levels of the variable
  - However, the sign of the coefficient would change.

**Source**: https://stats.idre.ucla.edu/spss/faq/coding-systems-for-categorical-variables-in-regression-analysis/

# SIMPLE REGRESSION CODING



| Level of race | New variable 1 (sx1) | New variable 2 (sx2) | New variable 3 (sx3) |
|---|---|---|---|
| 1 (White) | -1/4 | -1/4 | -1/4 |
| 2 (Black) | 3/4 | -1/4 | -1/4 |
| 3 (Hispanic) | -1/4 | 3/4 | -1/4 |
| 4 (Asian) | -1/4 | -1/4 | 3/4 |

- Level 1 is the reference level
- **sx1** compares level 2 to level 1
- **sx2** compares level 3 to level 1
- **sx3** compares level 4 to level 1.
- For s**x1** the coding is 3/4 for level 2, and -1/4 for all other levels.
- for s**x2** the coding is 3/4 for level 3, and -1/4 for all other levels,
- for s**x3** the coding is 3/4 for level 4, and -1/4 for all other levels.

**Source**: https://stats.idre.ucla.edu/stata/webbooks/reg/chapter5/regression-with-statachapter-5-additional-coding-systems-for-categorical-variables-in-regressionanalysis/

# General rule for simple coding



| Level of grouping variable | New variable 1 (x1) | New variable 2 (x2) | New variable 3 (x3) |
|---|---|---|---|
| Group 1 | -1 / k | -1 / k | -1 / k |
| Group 2 | (k-1) / k | -1 / k | -1 / k |
| Group 3 | -1 / k | (k-1) / k | -1 / k |
| Group k | -1 / k | -1 / k | (k-1) / k |

- It may not intuitive that this regression coding scheme yields these comparisons
  - Follow this general rule to obtain simple comparisons
- General rule for creating a simple coding scheme using regression coding
  - Where k is the number of levels of the categorical variable
  - In this example, k = 4

**Source**: https://stats.idre.ucla.edu/stata/webbooks/reg/chapter5/regression-with-statachapter-5-additional-coding-systems-for-categorical-variables-in-regressionanalysis/

# SIMPLE CODING REGRESSION MODEL INTERPRETATION (UNWEIGHTED)

The results of simple coding are very similar to dummy coding in that each level is compared to the reference level

| Source | SS | df | MS |
|---|---|---|---|
| Model | 20522.7022 | 3 | 6840.90072 |
| Residual | 310954.466 | 8,139 | 38.2054879 |
| Total | 331477.168 | 8,142 | 40.7120079 |

| | |
|---|---|
| Number of obs | = 8,143 |
| F(3, 8139) | = 179.06 |
| Prob > F | = 0.0000 |
| R-squared | = 0.0619 |
| Adj R-squared | = 0.0616 |
| Root MSE | = 6.1811 |

| bmi | Coefficient | Std. err. | t | P>\|t\| | [95% conf. | interval] |
|---|---|---|---|---|---|---|
| sx1 | 2.481632 | .1886871 | 13.15 | 0.000 | 2.111757 | 2.851507 |
| sx2 | 2.278446 | .173471 | 13.13 | 0.000 | 1.938398 | 2.618493 |
| sx3 | -1.817518 | .2093989 | -8.68 | 0.000 | -2.227994 | -1.407043 |
| _cons | 27.08866 | .0714918 | 378.91 | 0.000 | 26.94852 | 27.2288 |

| Race/ethnicity variable of person interviewed | Summary of B<br>Mean |
|---|---|
| White | 26.35302 |
| Black | 28.834652 |
| Hispanic | 28.631466 |
| Asian/PI | 24.535502 |
| Total | 27.264005 |

```
regress bmi sx1 sx2 sx3
```

- Level 1 (White) is the reference level
- x1 compares level 2 (Black) to level 1 = 28.83-26.35 =2.48
- X2 compares level 3 (Hispanic) to level 1 = 28.63-26.35= 2.28
- X3 compares level 4 (Asian) to level 1 = 24.54-26.35 = -1.81

# SIMPLE CODING REGRESSION MODEL INTERPRETATION (WEIGHTED)

The results of simple coding are very similar to dummy coding in that each level is compared to the reference level

```
Survey: Linear regression

Number of strata =   159                    Number of obs    =     8,143
Number of PSUs   = 8,143                    Population size = 5,988,401
                                            Design df        =     7,984
                                            F(3, 7982)       =     77.90
                                            Prob > F         =    0.0000
                                            R-squared        =    0.0533
```

| bmi | Coefficient | Linearized std. err. | t | P>\|t\| | [95% conf. | interval] |
|---|---|---|---|---|---|---|
| sx1 | 2.133186 | .3019247 | 7.07 | 0.000 | 1.541335 | 2.725037 |
| sx2 | 2.110904 | .2628895 | 8.03 | 0.000 | 1.595571 | 2.626236 |
| sx3 | -2.00541 | .2876277 | -6.97 | 0.000 | -2.569236 | -1.441585 |
| _cons | 26.88559 | .1071415 | 250.94 | 0.000 | 26.67556 | 27.09561 |

```
Survey: Mean estimation

Number of strata =   159
Number of PSUs   = 8,143
```

| | Mean |
|---|---|
| c.bmi@newrace | |
| White | 26.32592 |
| Black | 28.4591 |
| Hispanic | 28.43682 |
| Asian/PI | 24.32051 |

```
svy: regress bmi sx1 sx2 sx3
```

- Level 1 (White) is the reference level
- x1 compares level 2 (Black) to level 1 = 28.46-26.33 =2.13
- X2 compares level 3 (Hispanic) to level 1 = 28.44-26.33= 2.11
- X3 compares level 4 (Asian) to level 1 = 24.32-26.33 = -2.01

**Source**: https://stats.idre.ucla.edu/stata/webbooks/reg/chapter5/regression-with-statachapter-5-additional-coding-systems-for-categorical-variables-in-regressionanalysis/

# FORWARD DIFFERENCE CODING

- The mean of the dependent variable for one level of the categorical variable is compared to the mean of the dependent variable for the next (adjacent) level.

- First comparison compares the mean of **BMI** for level 1 with the mean of **BMI** for level 2 of **race**
  - White minus Black

- The second comparison compares the mean of **BMI** for level 2 minus level 3
  - Black minus Hispanic

- The third comparison compares the mean of **BMI** for level 3 minus level 4
  - Hispanic minus Asian



**Source**: https://stats.idre.ucla.edu/stata/webbooks/reg/chapter5/regression-with-statachapter-5-additional-coding-systems-for-categorical-variables-in-regressionanalysis/

# FORWARD DIFFERENCE REGRESSION CODING

- For the first comparison, the first and second levels are compared
  - **fx1** is coded 3/4 for level 1 and the other levels are coded -1/4
- For the second comparison where level 2 is compared with level 3
  - **fx2** is coded 1/2 1/2 -1/2 -1/2
- For the third comparison where level 3 is compared with level 4
  - **fx3** is coded 1/4 1/4 1/4 -3/4

| Level of race | New variable 1 (fx1) | New variable 2 (fx2) | New variable 3 (fx3) |
|---|---|---|---|
| | Level 1 v. Level 2 | Level 2 v. Level 3 | Level 3 v. Level 4 |
| 1 (White) | 3/4 | 1/2 | 1/4 |
| 2 (Black) | -1/4 | 1/2 | 1/4 |
| 3 (Hispanic) | -1/4 | -1/2 | 1/4 |
| 4 (Asian) | -1/4 | -1/2 | -3/4 |



**Source**: https://stats.idre.ucla.edu/stata/webbooks/reg/chapter5/regression-with-statachapter-5-additional-coding-systems-for-categorical-variables-in-regressionanalysis/

# GENERAL RULE FOR FORWARD DIFFERENCE CODING

- The general rule for forward difference regression coding scheme,
- Where k is the number of levels of the categorical variable (in this case k = 4)

| Groups | New variable 1 (x1) | New variable 2 (x2) | New variable 3 (x3) |
|---|---|---|---|
| | Group 1 v. Group 2 | Group 2 v. Group 3 | Group 3 v. Group 4 |
| Group 1 | (k-1)/k | (k-2)/k | (k-3)/k |
| Group 2 | -1/k | (k-2)/k | (k-3)/k |
| Group 3 | -1/k | -2/k | (k-3)/k |
| Group 4 | -1/k | -2/k | -3/k |



**Source**: https://stats.idre.ucla.edu/stata/webbooks/reg/chapter5/regression-with-statachapter-5-additional-coding-systems-for-categorical-variables-in-regressionanalysis/

# FORWARD DIFFERENCE CODING REGRESSION OUTPUT (UNWEIGHTED)

```
regress bmi fx1 fx2 fx3
```

| Source | SS | df | MS |
|---|---|---|---|
| Model | 20522.7022 | 3 | 6840.90072 |
| Residual | 310954.466 | 8,139 | 38.2054879 |
| Total | 331477.168 | 8,142 | 40.7120079 |

| | | | |
|---|---|---|---|
| Number of obs | = | 8,143 | |
| F(3, 8139) | = | 179.06 | |
| Prob > F | = | 0.0000 | |
| R-squared | = | 0.0619 | |
| Adj R-squared | = | 0.0616 | |
| Root MSE | = | 6.1811 | |

| bmi | Coefficient | Std. err. | t | P>\|t\| | [95% conf. | interval] |
|---|---|---|---|---|---|---|
| fx1 | -2.481632 | .1886871 | -13.15 | 0.000 | -2.851507 | -2.111757 |
| fx2 | .2031861 | .1947547 | 1.04 | 0.297 | -.1785828 | .584955 |
| fx3 | 4.095964 | .2148824 | 19.06 | 0.000 | 3.67474 | 4.517189 |
| _cons | 27.08866 | .0714918 | 378.91 | 0.000 | 26.94852 | 27.2288 |

# FORWARD DIFFERENCE CODING REGRESSION MODEL INTERPRETATION (UNWEIGHTED)

- With this coding system, adjacent levels of the categorical variable are compared
- You can see the regression coefficient for **fx1** is the mean of **BMI** for level 1 (White) minus the mean of **BMI** for level 2 (Black)
  - 26.35 – 28.83 = -2.48, which is statistically significant (p<0.001)
- The regression coefficient for **fx2** is the mean of **BMI** for level 2 (Black) minus the mean of **BMI** for level 3 (Hispanic)
  - The calculation of the contrast coefficient would be 28.83 – 28.63 = 0.2, which is not statistically significant (p=0.297)
- The regression coefficient for **fx3** is the mean of **BMI** for level 3 (Hispanic) minus the mean of **BMI** for level 4 (Asian)
  - 28.63- 24.54 =4.09, a statistically significant difference (p<0.001)

| Race/ethnicity variable of person interviewed | Summary of B<br>Mean |
|---|---|
| White | 26.35302 |
| Black | 28.834652 |
| Hispanic | 28.631466 |
| Asian/PI | 24.535502 |
| Total | 27.264005 |

| Source | SS | df | MS | | |
|---|---|---|---|---|---|
| Model | 20522.7022 | 3 | 6840.90072 | Number of obs | = 8,143 |
| Residual | 310954.466 | 8,139 | 38.2054879 | F(3, 8139) | = 179.06 |
| | | | | Prob > F | = 0.0000 |
| | | | | R-squared | = 0.0619 |
| | | | | Adj R-squared | = 0.0616 |
| Total | 331477.168 | 8,142 | 40.7120079 | Root MSE | = 6.1811 |

| bmi | Coefficient | Std. err. | t | P>\|t\| | [95% conf. | interval] |
|---|---|---|---|---|---|---|
| fx1 | -2.481632 | .1886871 | -13.15 | 0.000 | -2.851507 | -2.111757 |
| fx2 | .2031861 | .1947547 | 1.04 | 0.297 | -.1785828 | .584955 |
| fx3 | 4.095964 | .2148824 | 19.06 | 0.000 | 3.67474 | 4.517189 |
| _cons | 27.08866 | .0714918 | 378.91 | 0.000 | 26.94852 | 27.2288 |

# FORWARD DIFFERENCE CODING REGRESSION OUTPUT (WEIGHTED)

```
. svy: regress bmi fx1 fx2 fx3
```

```
Survey: Linear regression

Number of strata =   159                    Number of obs    =     8,143
Number of PSUs   = 8,143                    Population size  = 5,988,401
                                            Design df        =     7,984
                                            F(3, 7982)       =     77.90
                                            Prob > F         =    0.0000
                                            R-squared        =    0.0533
```

| bmi | Coefficient | Linearized std. err. | t | P>\|t\| | [95% conf. | interval] |
|---|---|---|---|---|---|---|
| fx1 | -2.133186 | .3019247 | -7.07 | 0.000 | -2.725037 | -1.541335 |
| fx2 | .0222823 | .3192645 | 0.07 | 0.944 | -.6035594 | .648124 |
| fx3 | 4.116314 | .3060669 | 13.45 | 0.000 | 3.516343 | 4.716285 |
| _cons | 26.88559 | .1071415 | 250.94 | 0.000 | 26.67556 | 27.09561 |

# FORWARD DIFFERENCE CODING REGRESSION MODEL INTERPRETATION (WEIGHTED)

- With this coding system, adjacent levels of the categorical variable are compared
- You can see the regression coefficient for **fx1** is the mean of **BMI** for level 1 (White) minus the mean of **BMI** for level 2 (Black)
  - 26.32 – 28.45 = -2.13, which is statistically significant (p<0.001)
- The regression coefficient for **fx2** is the mean of **BMI** for level 2 (Black) minus the mean of **BMI** for level 3 (Hispanic)
  - The calculation of the contrast coefficient would be 28.459 – 28.437= 0.022, which is not statistically significant (p=0.944)
- The regression coefficient for **fx3** is the mean of **BMI** for level 3 (Hispanic) minus the mean of **BMI** for level 4 (Asian)
  - 28.44- 24.32 =4.12, a statistically significant difference (p<0.001)

```
Survey: Mean estimation

Number of strata =   159
Number of PSUs   = 8,143
```

| | Mean |
|---|---|
| c.bmi@newrace | |
| White | 26.32592 |
| Black | 28.4591 |
| Hispanic | 28.43682 |
| Asian/PI | 24.32051 |

```
Survey: Linear regression

Number of strata =   159
Number of PSUs   = 8,143

                  Number of obs   =     8,143
                  Population size = 5,988,401
                  Design df       =     7,984
                  F(3, 7982)      =     77.90
                  Prob > F        =    0.0000
                  R-squared       =    0.0533
```

| bmi | Coefficient | Linearized std. err. | t | P>\|t\| | [95% conf. | interval] |
|---|---|---|---|---|---|---|
| fx1 | -2.133186 | .3019247 | -7.07 | 0.000 | -2.725037 | -1.541335 |
| fx2 | .0222823 | .3192645 | 0.07 | 0.944 | -.6035594 | .648124 |
| fx3 | 4.116314 | .3060669 | 13.45 | 0.000 | 3.516343 | 4.716285 |
| _cons | 26.88559 | .1071415 | 250.94 | 0.000 | 26.67556 | 27.09561 |

# BACKWARD DISTANCE CODING

- The mean of the dependent variable for one level of the categorical variable is compared to the mean of the dependent variable for the prior adjacent level.
- In our example, the first comparison variable compares the mean of **BMI** for level 2 (Black) with the mean of **BMI** for level 1 (White) of **race**
  - **bx1** is coded -3/4 for level 1 while the other levels are coded -1/4.
- The second comparison variable compares the mean of **BMI** for level 3 (Hispanic) minus level 2 (Black)
  - **bx2** is coded -1/2 -1/2 1/2 1/2
- The third comparison variable compares the mean of **BMI** for level 4 (Asian) minus level 3 (Hispanic)
  - **bx3** is coded -1/4 -1/4 -1/4 3/4.



**Source**: https://stats.idre.ucla.edu/stata/webbooks/reg/chapter5/regression-with-statachapter-5-additional-coding-systems-for-categorical-variables-in-regressionanalysis/

# BACKWARD DISTANCE REGRESSION CODING

| Race/Ethnicity | New variable 1 (bx1) | New variable 2 (bx2) | New variable 3 (bx3) |
|---|---|---|---|
| | **Black v. White** | **Hispanic v. Black** | **Asian v. Hispanic** |
| White | - 3/4 | -1/2 | -1/4 |
| Black | 1/4 | -1/2 | -1/4 |
| Hispanic | 1/4 | 1/2 | -1/4 |
| Asian | 1/4 | 1/2 | 3/4 |



**Source**: https://stats.idre.ucla.edu/stata/webbooks/reg/chapter5/regression-with-statachapter-5-additional-coding-systems-for-categorical-variables-in-regressionanalysis/

# GENERAL RULE FOR BACKWARD DISTANCE CODING

- The general rule for the backward distance regression coding scheme,
- Where k is the number of levels of the categorical variable (in this case, k = 4)

| Groups | New variable 1 (x1) | New variable 2 (x2) | New variable 3 (x3) |
|---|---|---|---|
| | Group 1 v. Group 2 | Group 2 v. Group 3 | Group 3 v. Group 4 |
| Group 1 | -(k-1)/k | -(k-2)/k | -(k-3)/k |
| Group 2 | 1/k | -(k-2)/k | -(k-3)/k |
| Group 3 | 1/k | 2/k | -(k-3)/k |
| Group 4 | 1/k | 2/k | 3/k |



**Source**: https://stats.idre.ucla.edu/stata/webbooks/reg/chapter5/regression-with-statachapter-5-additional-coding-systems-for-categorical-variables-in-regressionanalysis/

# BACKWARD DISTANCE CODING REGRESSION OUTPUT (UNWEIGHTED)

```
regress bmi bx1 bx2 bx3
```

| Source | SS | df | MS |
|---|---|---|---|
| Model | 20522.7022 | 3 | 6840.90072 |
| Residual | 310954.466 | 8,139 | 38.2054879 |
| Total | 331477.168 | 8,142 | 40.7120079 |

| | |
|---|---|
| Number of obs | = 8,143 |
| F(3, 8139) | = 179.06 |
| Prob > F | = 0.0000 |
| R-squared | = 0.0619 |
| Adj R-squared | = 0.0616 |
| Root MSE | = 6.1811 |

| bmi | Coefficient | Std. err. | t | P>\|t\| | [95% conf. | interval] |
|---|---|---|---|---|---|---|
| bx1 | 2.481632 | .1886871 | 13.15 | 0.000 | 2.111757 | 2.851507 |
| bx2 | -.2031861 | .1947547 | -1.04 | 0.297 | -.584955 | .1785828 |
| bx3 | -4.095964 | .2148824 | -19.06 | 0.000 | -4.517189 | -3.67474 |
| _cons | 27.08866 | .0714918 | 378.91 | 0.000 | 26.94852 | 27.2288 |

# BACKWARD DISTANCE CODING REGRESSION MODEL INTERPRETATION (UNWEIGHTED)

- With this coding system, adjacent levels of the categorical variable are compared
  - Each level compared to the prior level
- The regression coefficient for **bx1** is the mean of **BMI** for level 2 (Black) minus the mean of **BMI** for level 1 (White)
  - 28.83- 26.35 = 2.48; A statistically significant difference (p<0.001)
- The regression coefficient for **bx2** is the mean of **BMI** for level 3 (Hispanic) minus the mean of **BMI** for level 2 (Black)
  - 28.63 – 28.83 = -0.2; Not a statistically significant difference (p=0.297)
- The regression coefficient for **bx3** is the mean of **BMI** for level 4 (Asian) minus the mean of **BMI** for level 3 (Hispanic)
  - 24.54 – 28.63 = -4.09; A statistically significant difference (p<0.001)

| Race/ethnicity variable of person interviewed | Summary of B<br>Mean |
|---|---|
| White | 26.35302 |
| Black | 28.834652 |
| Hispanic | 28.631466 |
| Asian/PI | 24.535502 |
| Total | 27.264005 |

| Source | SS | df | MS | | |
|---|---|---|---|---|---|
| Model | 20522.7022 | 3 | 6840.90072 | Number of obs | = 8,143 |
| Residual | 310954.466 | 8,139 | 38.2054879 | F(3, 8139) | = 179.06 |
| | | | | Prob > F | = 0.0000 |
| Total | 331477.168 | 8,142 | 40.7120079 | R-squared | = 0.0619 |
| | | | | Adj R-squared | = 0.0616 |
| | | | | Root MSE | = 6.1811 |

| bmi | Coefficient | Std. err. | t | P>\|t\| | [95% conf. | interval] |
|---|---|---|---|---|---|---|
| bx1 | 2.481632 | .1886871 | 13.15 | 0.000 | 2.111757 | 2.851507 |
| bx2 | -.2031861 | .1947547 | -1.04 | 0.297 | -.584955 | .1785828 |
| bx3 | -4.095964 | .2148824 | -19.06 | 0.000 | -4.517189 | -3.67474 |
| _cons | 27.08866 | .0714918 | 378.91 | 0.000 | 26.94852 | 27.2288 |

# BACKWARD DISTANCE CODING REGRESSION OUTPUT (WEIGHTED)

```
svy: regress bmi bx1 bx2 bx3

Survey: Linear regression

Number of strata =   159                    Number of obs    =     8,143
Number of PSUs   = 8,143                    Population size = 5,988,401
                                            Design df        =     7,984
                                            F(3, 7982)       =     77.90
                                            Prob > F         =    0.0000
                                            R-squared        =    0.0533
```

| bmi | Coefficient | Linearized std. err. | t | P>\|t\| | [95% conf. | interval] |
|---|---|---|---|---|---|---|
| bx1 | 2.133186 | .3019247 | 7.07 | 0.000 | 1.541335 | 2.725037 |
| bx2 | -.0222823 | .3192645 | -0.07 | 0.944 | -.648124 | .6035594 |
| bx3 | -4.116314 | .3060669 | -13.45 | 0.000 | -4.716285 | -3.516343 |
| _cons | 26.88559 | .1071415 | 250.94 | 0.000 | 26.67556 | 27.09561 |

# BACKWARD DISTANCE CODING REGRESSION MODEL INTERPRETATION (WEIGHTED)

- With this coding system, adjacent levels of the categorical variable are compared
  - Each level compared to the prior level
- The regression coefficient for **bx1** is the mean of **BMI** for level 2 (Black) minus the mean of **BMI** for level 1 (White)
  - 28.46- 26.33 = 2.13A statistically significant difference (p<0.001)
- The regression coefficient for **bx2** is the mean of **BMI** for level 3 (Hispanic) minus the mean of **BMI** for level 2 (Black)
  - 28.44 – 28.46 = -0.02; Not a statistically significant difference (p=0.944)
- The regression coefficient for **bx3** is the mean of **BMI** for level 4 (Asian) minus the mean of **BMI** for level 3 (Hispanic)
  - 24.32 – 28.44 = -4.12; A statistically significant difference (p<0.001)

| | Mean |
|---|---|
| c.bmi@newrace | |
| White | 26.32592 |
| Black | 28.4591 |
| Hispanic | 28.43682 |
| Asian/PI | 24.32051 |

Survey: Linear regression

Number of strata = 159
Number of PSUs = 8,143

Number of obs = 8,143
Population size = 5,988,401
Design df = 7,984
F(3, 7982) = 77.90
Prob > F = 0.0000
R-squared = 0.0533

| bmi | Coefficient | Linearized std. err. | t | P>\|t\| | [95% conf. | interval] |
|---|---|---|---|---|---|---|
| bx1 | 2.133186 | .3019247 | 7.07 | 0.000 | 1.541335 | 2.725037 |
| bx2 | -.0222823 | .3192645 | -0.07 | 0.944 | -.648124 | .6035594 |
| bx3 | -4.116314 | .3060669 | -13.45 | 0.000 | -4.716285 | -3.516343 |
| _cons | 26.88559 | .1071415 | 250.94 | 0.000 | 26.67556 | 27.09561 |

# DEVIATION CODING

- This coding system compares the mean of the dependent variable for a given level to the mean of the dependent variable for the all levels of the variable.
  - In our example, the first comparison variable compares level 2 (Blacks) to all levels of **race**
  - The second new variable compares level 3 (Hispanics) to all levels of **race**
  - The third comparison variable compares level 4 (Asians) to all levels of race

- The regression coding is accomplished by assigning
  - 1 to level 2 for the first comparison (because level 2 is the level to be compared to all)
  - 1 to level 3 for the second comparison (because level 3 is to be compared to all)
  - 1 to level 4 for the third comparison (because level 4 is to be compared to all)
  - Note that a -1 is assigned to level 1 for all three comparisons (because it is the level that is never compared to the other levels) and all other values are assigned a 0



**Source**: https://stats.idre.ucla.edu/stata/webbooks/reg/chapter5/regression-with-statachapter-5-additional-coding-systems-for-categorical-variables-in-regressionanalysis/

# DEVIATION REGRESSION CODING

| Race/Ethnicity | New variable 1 (dx1) | New variable 2 (dx2) | New variable 3 (dx3) |
|---|---|---|---|
| | *Black v. Mean* | *Hispanic v. Mean* | *Asian v. Mean* |
| White | -1 | -1 | -1 |
| Black | 1 | 0 | 0 |
| Hispanic | 0 | 1 | 0 |
| Asian | 0 | 0 | 1 |



**Source**: https://stats.idre.ucla.edu/stata/webbooks/reg/chapter5/regression-with-statachapter-5-additional-coding-systems-for-categorical-variables-in-regressionanalysis/

# DEVIATION CODING REGRESSION OUTPUT (UNWEIGHTED)

```
regress bmi dx1 dx2 dx3
```

| Source | SS | df | MS |
|---|---|---|---|
| Model | 20522.7022 | 3 | 6840.90072 |
| Residual | 310954.466 | 8,139 | 38.2054879 |
| Total | 331477.168 | 8,142 | 40.7120079 |

| | |
|---|---|
| Number of obs | = 8,143 |
| F(3, 8139) | = 179.06 |
| Prob > F | = 0.0000 |
| R-squared | = 0.0619 |
| Adj R-squared | = 0.0616 |
| Root MSE | = 6.1811 |

| bmi | Coefficient | Std. err. | t | P>\|t\| | [95% conf. | interval] |
|---|---|---|---|---|---|---|
| dx1 | 1.745992 | .1263767 | 13.82 | 0.000 | 1.498261 | 1.993723 |
| dx2 | 1.542806 | .1149598 | 13.42 | 0.000 | 1.317455 | 1.768157 |
| dx3 | -2.553158 | .141752 | -18.01 | 0.000 | -2.831028 | -2.275288 |
| _cons | 27.08866 | .0714918 | 378.91 | 0.000 | 26.94852 | 27.2288 |

# DEVIATION CODING REGRESSION COEFFICIENTS (UNWEIGHTED)

- The regression coefficient for **dx1** is the mean for level 2 (Black) minus the grand mean.
  - However, this grand mean is not the overall mean of the dependent variable that you would get from taking the mean (27.24) of all of the observations
  - Rather, it is the mean of means of the dependent variable at each level of the categorical variable
  - (26.35+ 28.83 + 28.63 + 24.54) / 4 = 27.088
  - The regression coefficient is then 28.83 – 27.088 = 1.74; this difference is statistically significant (p<0.001)

| Variable | Obs | Mean | Std. dev. | Min | Max |
|---|---|---|---|---|---|
| bmi | 8,423 | 27.23568 | 6.352999 | 7.933049 | 82.16812 |

| Race/ethnicity variable of person interviewed | Summary of B<br>Mean |
|---|---|
| White | 26.35302 |
| Black | 28.834652 |
| Hispanic | 28.631466 |
| Asian/PI | 24.535502 |
| Total | 27.264005 |

# DEVIATION CODING REGRESSION INTERPRETATION (UNWEIGHTED)

- The coefficient for **dx2** is the mean for level 3 (Hispanic) of race minus the overall mean
  - 28.63- 27.088 = 1.54; this difference is statistically significant (p<0.001)
- The coefficient for **dx3** is the mean for level 4 (Asian) of race minus the overall mean
  - 24.54 - 27.088 = -2.55; this difference is statistically significant (p<0.001)

```
. regress bmi dx1 dx2 dx3
```

| Source | SS | df | MS |
|---|---|---|---|
| Model | 20522.7022 | 3 | 6840.90072 |
| Residual | 310954.466 | 8,139 | 38.2054879 |
| Total | 331477.168 | 8,142 | 40.7120079 |

| | |
|---|---|
| Number of obs | = 8,143 |
| F(3, 8139) | = 179.06 |
| Prob > F | = 0.0000 |
| R-squared | = 0.0619 |
| Adj R-squared | = 0.0616 |
| Root MSE | = 6.1811 |

| bmi | Coefficient | Std. err. | t | P>\|t\| | [95% conf. | interval] |
|---|---|---|---|---|---|---|
| dx1 | 1.745992 | .1263767 | 13.82 | 0.000 | 1.498261 | 1.993723 |
| dx2 | 1.542806 | .1149598 | 13.42 | 0.000 | 1.317455 | 1.768157 |
| dx3 | -2.553158 | .141752 | -18.01 | 0.000 | -2.831028 | -2.275288 |
| _cons | 27.08866 | .0714918 | 378.91 | 0.000 | 26.94852 | 27.2288 |

# DEVIATION CODING REGRESSION OUTPUT (WEIGHTED)

```
. svy: regress bmi dx1 dx2 dx3
(running regress on estimation sample)

Survey: Linear regression

Number of strata =   159                  Number of obs    =     8,143
Number of PSUs   = 8,143                  Population size  = 5,988,401
                                          Design df        =     7,984
                                          F(3, 7982)       =     77.90
                                          Prob > F         =    0.0000
                                          R-squared        =    0.0533
```

| bmi | Coefficient | Linearized std. err. | t | P>\|t\| | [95% conf. | interval] |
|---|---|---|---|---|---|---|
| dx1 | 1.573516 | .2060685 | 7.64 | 0.000 | 1.169568 | 1.977464 |
| dx2 | 1.551234 | .1776651 | 8.73 | 0.000 | 1.202964 | 1.899504 |
| dx3 | -2.56508 | .195688 | -13.11 | 0.000 | -2.94868 | -2.181481 |
| _cons | 26.88559 | .1071415 | 250.94 | 0.000 | 26.67556 | 27.09561 |

# DEVIATION CODING REGRESSION COEFFICIENTS (WEIGHTED)

- The regression coefficient for **dxl** is the mean for level 2 (Black) minus the grand mean.
  - However, this grand mean is not the overall mean of the dependent variable that you would get from taking the mean (27.09) of all of the observations
  - Rather, it is the mean of means of the dependent variable at each level of the categorical variable
  - (26.33+ 28.46 + 28.44 + 24.32) / 4 = 26.89
  - The regression coefficient is then 28.46 – 26.89 = 1.57

```
Survey: Mean estimation

Number of strata =   159          Number of obs    =     8,423
Number of PSUs   = 8,423          Population size = 6,137,719
                                  Design df       =     8,264
```

| | Linearized Mean | std. err. | [95% conf. | interval] |
|---|---|---|---|---|
| bmi | 27.09084 | .1046812 | 26.88564 | 27.29604 |

| | Mean |
|---|---|
| c.bmi@newrace | |
| White | 26.32592 |
| Black | 28.4591 |
| Hispanic | 28.43682 |
| Asian/PI | 24.32051 |

# DEVIATION CODING REGRESSION INTERPRETATION (WEIGHTED)

- The coefficient for **dx2** is the mean for level 3 (Hispanic) of race minus the overall mean
  - 28.44- 26.89 = 1.55
- The coefficient for **dx3** is the mean for level 4 (Asian) of race minus the overall mean
  - 24.32 - 26.89 = -2.57

```
. svy: regress bmi dx1 dx2 dx3
(running regress on estimation sample)

Survey: Linear regression

Number of strata =   159                  Number of obs    =     8,143
Number of PSUs   = 8,143                  Population size  = 5,988,401
                                          Design df        =     7,984
                                          F(3, 7982)       =     77.90
                                          Prob > F         =    0.0000
                                          R-squared        =    0.0533
```

| bmi | Linearized Coefficient | std. err. | t | P>\|t\| | [95% conf. | interval] |
|---|---|---|---|---|---|---|
| dx1 | 1.573516 | .2060685 | 7.64 | 0.000 | 1.169568 | 1.977464 |
| dx2 | 1.551234 | .1776651 | 8.73 | 0.000 | 1.202964 | 1.899504 |
| dx3 | -2.56508 | .195688 | -13.11 | 0.000 | -2.94868 | -2.181481 |
| _cons | 26.88559 | .1071415 | 250.94 | 0.000 | 26.67556 | 27.09561 |

# STATISTICAL ANALYST DEFINED CODING

- You can create your own regression coding system based on the comparisons you are interested in making
- Suppose we want to make the following three comparisons:
  - 1) level 1 (White) to level 4 (Asian)
    - In order to compare level 1 to level 4, we use the contrast coefficients 1 0 0 -1
  - 2) level 2 (Black) to levels 1 and 4 (White and Asian)
    - To compare level 2 to levels 1 and 4 we use the contrast coefficients -1/2 1 0 -1/2
  - 3) levels 2 and 3 (Black and Hispanic) to levels 1 and 4 (White and Asian)
    - To compare levels 2 and 3 with levels 1 and 4 we use the contrast coefficients 1/2 -1/2 -1/2 1/2



**Source**: https://stats.idre.ucla.edu/stata/webbooks/reg/chapter5/regression-with-statachapter-5-additional-coding-systems-for-categorical-variables-in-regressionanalysis/

# KEY POINTS: CONTRASTS

- For the first contrast, we are comparing level 1 to level 4, and the contrast coefficients are 1 0 0 -1
- The levels associated with the contrast coefficients with opposite signs are being compared
- The mean of the dependent variable is multiplied by the contrast coefficient
- Hence, levels 2 and 3 are not involved in the comparison
  - They are multiplied by zero and "dropped out"
- You will also notice that the contrast coefficients sum to zero
  - This is necessary
  - If the contrast coefficients do not sum to zero, the contrast is not estimable
- Which level of the categorical variable is assigned a positive or negative value is not terribly important
  - 1 0 -1 0 is the same as -1 0 1 0 in that both of these codings compare the first and the third levels of the variable
  - However, the sign of the regression coefficient would change



**Source**: https://stats.idre.ucla.edu/stata/webbooks/reg/chapter5/regression-with-statachapter-5-additional-coding-systems-for-categorical-variables-in-regressionanalysis/

# DETERMINE THE REGRESSION CODING SYSTEM

- For methods 1 and 2 it was quite easy to translate the comparisons we wanted to make into contrast codings, but it is not as easy to translate the comparisons we want into a regression coding scheme.
- If we know the contrast coding system, then we can convert that into a regression coding system
- We place the three contrast codings we want into the matrix c and then perform a set of matrix operations on c, yielding the matrix x.
  - `matrix input c = (1 0 0 -1 \ -.5 1 0 -.5 \ .5 -.5 -.5 .5)`
  - `matrix list c`
  - `matrix x = c'*inv(c*c')`
  - `matrix list x`

```
c[3,4]
     c1   c2   c3   c4
r1    1    0    0   -1
r2  -.5    1    0  -.5
r3   .5  -.5  -.5   .5
```

```
x[4,3]
      r1   r2    r3
c1    .5    0    .5
c2     0    1    .5
c3     0   -1  -1.5
c4  -.5     0    .5
```



**Source**: https://stats.idre.ucla.edu/stata/webbooks/reg/chapter5/regression-with-statachapter-5-additional-coding-systems-for-categorical-variables-in-regressionanalysis/

# STATISTICAL ANALYST DEFINED CODING REGRESSION OUTPUT (UNWEIGHTED)

```
regress bmi ax1 ax2 ax3
```

| Source | SS | df | MS |
|---|---|---|---|
| Model | 20522.7022 | 3 | 6840.90072 |
| Residual | 310954.466 | 8,139 | 38.2054879 |
| Total | 331477.168 | 8,142 | 40.7120079 |

| | | | |
|---|---|---|---|
| Number of obs | = | 8,143 |
| F(3, 8139) | = | 179.06 |
| Prob > F | = | 0.0000 |
| R-squared | = | 0.0619 |
| Adj R-squared | = | 0.0616 |
| Root MSE | = | 6.1811 |

| bmi | Coefficient | Std. err. | t | P>\|t\| | [95% conf. | interval] |
|---|---|---|---|---|---|---|
| ax1 | 1.817518 | .2093989 | 8.68 | 0.000 | 1.407043 | 2.227994 |
| ax2 | 3.390391 | .1807816 | 18.75 | 0.000 | 3.036013 | 3.744769 |
| ax3 | -3.288798 | .1429836 | -23.00 | 0.000 | -3.569083 | -3.008514 |
| _cons | 27.08866 | .0714918 | 378.91 | 0.000 | 26.94852 | 27.2288 |

# STATISTICAL ANALYST DEFINED CODING INTERPRETATION (UNWEIGHTED)

- The coefficient for **ax1 (-1.82)** corresponds to the first contrast comparing level 1 (White) to level 4 (Asian) of **race**
  - The coefficient is the mean of level 1 (White) of **BMI** minus the mean for level 4 (Asian) of **BMI**
  - 26.35-24.54 = 1.8
  - p-value<0.001 (statistically significant)

| Race/ethnicity variable of person interviewed | Summary of B<br>Mean |
|---|---|
| White | 26.35302 |
| Black | 28.834652 |
| Hispanic | 28.631466 |
| Asian/PI | 24.535502 |
| Total | 27.264005 |

- The coefficient for **ax2** is 3.39 which is the mean of level 2 (Black) minus the mean of level 1 and level 4 (White and Asian)
  - $28.83 - \frac{(26.35+24.54)}{2}$= 28.83 -25.445=3.39
  - This difference is statistically significant, p<0.001

- The coefficient for **ax3** is -3.29 which is the mean of levels 1 and 4 (White and Asian) minus the mean of levels 2 and 3 (Black and Hispanic )=

$$\frac{(26.35 + 24.54)}{2} - \frac{(28.83 + 28.63)}{2}$$

- 25.445-28.73 = -3.39
  - This contrast is statistically significant, p<0.001

# STATISTICAL ANALYST DEFINED CODING REGRESSION OUTPUT (WEIGHTED)

```
svy: regress bmi ax1 ax2 ax3
```

```
Survey: Linear regression

Number of strata =   159                    Number of obs    =     8,143
Number of PSUs   = 8,143                    Population size = 5,988,401
                                            Design df        =     7,984
                                            F(3, 7982)       =     77.90
                                            Prob > F         =    0.0000
                                            R-squared        =    0.0533
```

| bmi | Coefficient | Linearized std. err. | t | P>\|t\| | [95% conf. | interval] |
|---|---|---|---|---|---|---|
| ax1 | 2.00541 | .2876277 | 6.97 | 0.000 | 1.441585 | 2.569236 |
| ax2 | 3.135891 | .2872458 | 10.92 | 0.000 | 2.572814 | 3.698968 |
| ax3 | -3.12475 | .2147752 | -14.55 | 0.000 | -3.545765 | -2.703734 |
| _cons | 26.88559 | .1071415 | 250.94 | 0.000 | 26.67556 | 27.09561 |

# STATISTICAL ANALYST DEFINED CODING INTERPRETATION (WEIGHTED)

- The coefficient for **ax1 (2.0)** corresponds to the first contrast comparing level 1 (White) to level 4 (Asian) of **race**
  - The coefficient is the mean of level 1 (White) of **BMI** minus the mean for level 4 (Asian) of **BMI**
  - 26.33-24.32 = 2.0
  - p-value<0.001 (statistically significant)

| | Mean |
|---|---|
| c.bmi@newrace | |
| White | 26.32592 |
| Black | 28.4591 |
| Hispanic | 28.43682 |
| Asian/PI | 24.32051 |

- The coefficient for **ax2** is 3.14 which is the mean of level 2 (Black) minus the mean of level 1 and level 4 (White and Asian)
  - $28.46 - \frac{(26.33+24.32)}{2}$= 28.46 -25.325=3.14
  - This difference is statistically significant, p<0.001

- The coefficient for **ax3** is -3.12 which is the mean of levels 1 and 4 (White and Asian) minus the mean of levels 2 and 3 (Black and Hispanic )=

$$\frac{(26.33 + 24.32)}{2} - \frac{(28.46 + 28.44)}{2}$$

- 25.33-28.45 = -3.12
  - This contrast is statistically significant, p<0.001

# USING FRACTIONS IN CONTRASTS

- In the contrast coefficients for the second and third comparisons, notice that in both cases we use fractions that sum to one (or minus one)
  - Note they do not have to sum to one (or minus one)

- You may wonder why we would use fractions like -1/2 1 0 -1/2 instead of whole numbers such as -1 2 0 -1.
- While -1/2 1 0 -1/2 and -1 2 0 -1 both compare level 2 with levels 1 and 4
  - Both give you the same t-value and p-value for the regression coefficient
  - The regression coefficients would be different, as would their interpretation

- The coefficient for the -1/2 1 0 -1/2 contrast is the mean of level 2 minus the mean of the means for levels 1 and 4
  - $28.46 - \frac{(26.33+24.32)}{2}$= 28.46 -25.325=3.14

| | Mean |
|---|---|
| c.bmi@newrace | |
| White | 26.32592 |
| Black | 28.4591 |
| Hispanic | 28.43682 |
| Asian/PI | 24.32051 |



**Source**: https://stats.idre.ucla.edu/stata/webbooks/reg/chapter5/regression-with-statachapter-5-additional-coding-systems-for-categorical-variables-in-regressionanalysis/

# USING FRACTIONS IN CONTRASTS CONTINUED

- Alternatively, you can multiply the contrasts by the mean of the dependent variable for each level of the categorical variable:
  - -1/2*26.33+ 1*28.46 + 0*28.44 + -1/2*24.32 =-13.165+28.46-12.16=3.14

- These are equivalent ways of thinking about how the contrast coefficient is calculated and give the same value

- By comparison, the coefficient for the -1 2 0 -1 contrast is two times the mean for level 2 minus the means of the dependent variable for levels 1 and 4
  - 2*28.46 - (26.33 + 24.32) =56.92-50.65=6.27
  - Which is the same as -1* 26.33 + 2*28.46 + 0*28.44- 1* 24.32
  - =-26.33+56.92-24.32=6.27

- Note that the regression coefficient using the contrast coefficients -1 2 0 -1 is twice the regression coefficient obtained when -1/2 1 0 -1/2 is used

| | Mean |
|---|---|
| c.bmi@newrace | |
| White | 26.32592 |
| Black | 28.4591 |
| Hispanic | 28.43682 |
| Asian/PI | 24.32051 |



**Source**: https://stats.idre.ucla.edu/stata/webbooks/reg/chapter5/regression-with-statachapter-5-additional-coding-systems-for-categorical-variables-in-regressionanalysis/

# THINK CRITICALLY ABOUT MODELING RACE IN REGRESSION

- It is insufficient to control for race in a regression model without thinking critically about what that means
  - Race is not a modifiable factor

- How will race/ethnicity be measured and analyzed?
  - Use appropriate scales validated in your population of interested

- How is race being used in the presentation of the results?
  - Is race a proxy for something else?

- Is it racism and not race?
  - Think about how racism is operating and could impact the results of your research study
  - If race is social and not biological is race really operating at the individual level?

NYU

SCHOOL
OF GLOBAL
PUBLIC HEALTH

# QUESTIONS?


Melody S. Goodman, PhD
Associate Dean for Research
Associate Professor of Biostatistics
NYU School of Global Public Health

melody.goodman@nyu.edu

@goodmanthebrain