

450A Political Methodology I

Lecture 5: Regression with Two Regressors

3. Dummy Variables

Stanford University

Dummy Variables

Two Categories

Multiple Categories

Two Dummy Variables and Their Interaction

Why Dummy Variables?

- A **dummy variable** (a.k.a. indicator variable, binary variable, etc.) is a variable that is coded 1 or 0 only.
- We use dummy variables in regression to represent qualitative information through **categorical variables** such as different subgroups of the sample (e.g. regions, old and young respondents, etc.)
- By including dummy variables into our regression function, we can easily obtain the **conditional mean of the outcome variable for each category**.
 - E.g. does average income vary by region?
- Dummy variables are also used to examine conditional hypotheses via **interaction terms**
 - E.g. does the effect of education differ by gender?

Dummy Variables

Two Categories

Multiple Categories

Two Dummy Variables and Their Interaction

How Can I Use a Dummy Variable?

- Consider the easiest case with two categories. In the 2016 referendum on leaving the European Union (Brexit), votes were counted in local areas. Counting area i of Great Britain lies in:
 $X_i \in \{\text{England \& Wales, Scotland}\}$
- For this we use a single dummy variable which is coded like:

$$D_i = \begin{cases} 1 & \text{if area } i \text{ is in Scotland} \\ 0 & \text{if area } i \text{ is in England or Wales} \end{cases}$$

- Hint: Informative variable names help (e.g. call it Scotland)
- Let's regress the Leave share (% of votes for Leave) on this dummy variable and a constant:
 $Y = \beta_0 + \beta_1 D + u$

Example: Leave Share on Scotland

R Code

```
> summary(lm(Leave ~ Scotland, data = D))

Call:
lm(formula = Leave ~ Scotland, data = D)

Residuals:
    Min       1Q   Median       3Q      Max
-33.048  -3.941   0.857   6.019  21.132

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  54.4285     0.5104 106.647  <2e-16 ***
Scotland    -15.2916     1.7587  -8.695  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.521 on 378 degrees of freedom
Multiple R-squared:  0.1667,    Adjusted R-squared:  0.1645 
F-statistic: 75.6 on 1 and 378 DF,  p-value: < 2.2e-16
```

Source: Electoral Commission, EU referendum results by counting area (2016): the 380 counting areas of Great Britain (England 326, Wales 22, Scotland 32); Northern Ireland and Gibraltar left out. Leave = % of valid votes.

Example: Leave Share on Scotland

R Code

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	54.4285	0.5104	106.647	<2e-16 ***
Scotland	-15.2916	1.7587	-8.695	<2e-16 ***

R Code

```
> leave.ew <- D$Leave[D$Scotland == 0]
> summary(leave.ew)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 21.38  50.35   55.28   54.43  60.87   75.56

> leave.sc <- D$Leave[D$Scotland == 1]
> summary(leave.sc)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 25.56  36.68   39.83   39.14  43.15   49.87
```

So this is just like a pooled-variance (equal variances) two-sample t-test for a difference in means! (`t.test(Leave ~ Scotland, data = D, var.equal = TRUE)` gives $t = 8.69$. It is the same test, with the groups subtracted the other way round.)

Example: Leave Share on Scotland

R Code

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	54.4285	0.5104	106.647	<2e-16 ***
Scotland	-15.2916	1.7587	-8.695	<2e-16 ***

R Code

```
> leave.ew <- D$Leave[D$Scotland == 0]
> summary(leave.ew)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 21.38  50.35   55.28   54.43  60.87   75.56

> leave.sc <- D$Leave[D$Scotland == 1]
> summary(leave.sc)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 25.56  36.68   39.83   39.14  43.15   49.87
```

Example: Leave Share on Scotland

R Code

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	54.4285	0.5104	106.647	<2e-16 ***
Scotland	-15.2916	1.7587	-8.695	<2e-16 ***

R Code

```
> leave.ew <- D$Leave[D$Scotland == 0]
> summary(leave.ew)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
21.38  50.35   55.28   54.43  60.87   75.56

> leave.sc <- D$Leave[D$Scotland == 1]
> summary(leave.sc)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
25.56  36.68   39.83   39.14  43.15   49.87
```

Regression with Dummy Variable for Two Categories

So our prediction equation is: $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 D_i$. What is the expected difference in the average Leave share between areas in **Scotland** and areas in **England & Wales**?

- $\hat{E}[\text{Leave}|\text{Scotland}] = \hat{E}[Y|D_i = 1] = \hat{\beta}_0 + \hat{\beta}_1$
- $\hat{E}[\text{Leave}|\text{England \& Wales}] = \hat{E}[Y|D_i = 0] = \hat{\beta}_0 + \hat{\beta}_1 \cdot 0 = \hat{\beta}_0$
- The expected difference is thus: $\hat{E}[Y|\text{Scotland}] - \hat{E}[Y|\text{England \& Wales}] = \hat{\beta}_0 + \hat{\beta}_1 - \hat{\beta}_0 = \hat{\beta}_1$
- Notice that dummy variables act like “switches” that turn various parameters on and off in an equation.
- What changes when we control for additional X variables (like the share of residents with no qualifications)?

Leave Share on Scotland and Education

R Code

```
> summary(lm(Leave ~ Scotland + NoQual, data = D))

Call:
lm(formula = Leave ~ Scotland + NoQual, data = D)

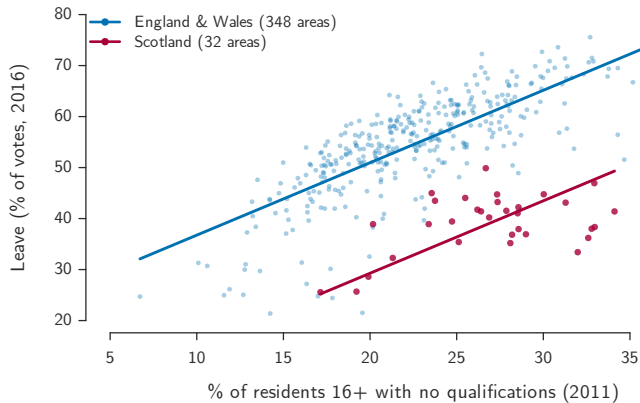
Residuals:
    Min       1Q   Median       3Q      Max
-28.811  -2.753   1.171   4.267  12.163

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  22.55043    1.50680   14.97  <2e-16 ***
Scotland    -21.67527    1.20997  -17.91  <2e-16 ***
NoQual       1.42020    0.06539   21.72  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

NoQual is the % of residents aged 16+ with no qualifications. Now we have the same slope for the partial relationship between the Leave share and NoQual, but two different intercepts for Scotland and for England & Wales.

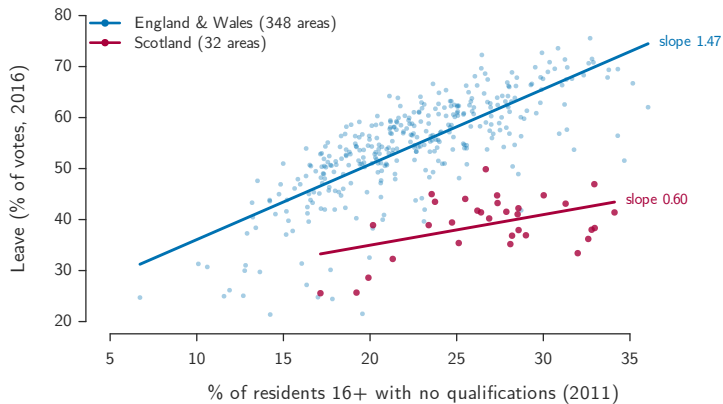
Source: NoQual: 2011 Census, table KS501UK (ONS, via Nomis).

Same Slope, Two Intercepts? The Data Disagree



The model on the last slide draws two parallel lines. Both have slope 1.42, and Scotland's line is 21.68 points lower.

Same Slope, Two Intercepts? The Data Disagree



Fitted to each group on its own, the lines are not parallel. Scotland's slope (0.60) is less than half that of England & Wales (1.47). The difference is about -0.9 ($t = -3.3$).

Letting the slope differ by group is an **interaction**, the topic of the next lecture (5-4).

Dummy Variables

Two Categories

Multiple Categories

Two Dummy Variables and Their Interaction

Dummy Variables for Multiple Categories

- More generally, let's say X measures which of m categories each unit i belongs to. E.g. the part of the country or the region of counting area i is given by:
 - $X_i \in \{\text{England \& Wales, Scotland}\}$ so $m = 2$
 - $X_i \in \{\text{London, South East, \dots, North West, Wales, Scotland}\}$, the 11 regions of Great Britain, so $m = 11$
- To incorporate this information into our regression function we usually create $m - 1$ dummy variables, one for each of the $m - 1$ categories.
- Why not all m ? Including all m category indicators together with an intercept would violate the no perfect collinearity assumption:

$$D_m = 1 - (D_1 + \dots + D_{m-1})$$

- The omitted category is our **baseline case** (also called a **reference category**) against which we compare the conditional means of Y for the other $m - 1$ categories.

Example: Regions of Great Britain

- Consider the case of our “polytomous” variable region with $m = 11$: London and the 8 other regions of England, Wales, and Scotland
- This eleven-category classification can be represented in the regression equation by introducing $m - 1 = 10$ dummy regressors, with London as the baseline:

Category	D_1	D_2	$\dots$	D_9	D_{10}
South East	1	0	$\dots$	0	0
South West	0	1	$\dots$	0	0
$\vdots$			$\ddots$		
Wales	0	0	$\dots$	1	0
Scotland	0	0	$\dots$	0	1
London	0	0	$\dots$	0	0

- Our regression equation is:

$$Y = \beta_0 + \beta_1 D_1 + \beta_2 D_2 + \dots + \beta_9 D_9 + \beta_{10} D_{10} + u$$

- Let's use the Leave share again as our DV and run the numbers

Example: Leave Share on Regions

R Code

```
> summary(lm(Leave ~ SouthEast + SouthWest + East + EastMidlands + WestMidlands
+ Yorkshire + NorthEast + NorthWest + Wales + Scotland, data = D))

~~~~~

Coefficients:
(Intercept) 39.09152      1.37272  28.477 < 2e-16 ***
SouthEast    13.07848      1.67704   7.799 6.48e-14 ***
SouthWest    14.59281      1.88812   7.729 1.04e-13 ***
East         17.87061      1.79093   9.978 < 2e-16 ***
EastMidlands 20.48298      1.85444  11.045 < 2e-16 ***
WestMidlands 21.22315      1.98926  10.669 < 2e-16 ***
Yorkshire    19.55848      2.20125   8.885 < 2e-16 ***
NorthEast    20.38765      2.65826   7.670 1.55e-13 ***
NorthWest    16.82361      1.86516   9.020 < 2e-16 ***
Wales        14.25621      2.17046   6.568 1.73e-10 ***
Scotland     0.04536      1.95643   0.023  0.982

---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Residual standard error: 7.886 on 369 degrees of freedom
Multiple R-squared:  0.4419,    Adjusted R-squared:  0.4268 
F-statistic: 29.22 on 10 and 369 DF,  p-value: < 2.2e-16
```

Worksheet Q1

Suppose X_i is the region of counting area i : London, 8 other English regions, Wales, or Scotland. Estimate

$$Y_i = \beta_0 + \beta_1 D_{SE,i} + \beta_2 D_{SW,i} + \cdots + \beta_9 D_{Wales,i} + \beta_{10} D_{Scot,i} + u_i,$$

with London omitted.

- 1 Interpret β_0
- 2 Interpret each β_j , $j \in \{1, \dots, 10\}$.
- 3 Derive the estimand and an OLS t -test for $H_0 : E[Y|Wales] = E[Y|Scotland]$. Express the test as a linear restriction $R\beta = \mathbf{0}$ and give the variance formula used for its standard error.

Dealing with a Categorical Variable in R

- In fact, R automatically expands a factor (categorical) variable with m levels into $m - 1$ dummy variables:

```
> levels(D$Region)
[1] "East"           "East Midlands"
[3] "London"         "North East"
[5] "North West"     "Scotland"
[7] "South East"     "South West"
[9] "Wales"          "West Midlands"
[11] "Yorkshire and The Humber"
```

R Code

```
> summary(lm(Leave ~ Region, data = D))
~~~~~
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    56.962     1.150   49.522 < 2e-16 ***
RegionEast Midlands  2.612     1.696    1.540  0.12442
RegionLondon   -17.871     1.791   -9.978 < 2e-16 ***
RegionNorth East  2.517     2.550    0.987  0.32435
RegionNorth West -1.047     1.708   -0.613  0.54027
RegionScotland  -17.825     1.807   -9.863 < 2e-16 ***
RegionSouth East -4.792     1.500   -3.194  0.00152 **
RegionSouth West -3.278     1.733   -1.891  0.05937 .
RegionWales     -3.614     2.037   -1.774  0.07683 .
RegionWest Midlands  3.353     1.843    1.819  0.06968 .
RegionYorkshire and The Humber 1.688     2.070    0.815  0.41533
```

R takes the first level in alphabetical order, East, as the baseline.

Dealing with a Categorical Variable in R

- You can change the baseline category by the `relevel()` function:

```
R Code
> D$Region <- relevel(D$Region, ref = "London")
> summary(lm(Leave ~ Region, data = D))
~~~~~
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   39.09152    1.37272  28.477 < 2e-16 ***
RegionEast    17.87061    1.79093   9.978 < 2e-16 ***
RegionEast Midlands 20.48298    1.85444  11.045 < 2e-16 ***
RegionNorth East 20.38765    2.65826   7.670 1.55e-13 ***
RegionNorth West 16.82361    1.86516   9.020 < 2e-16 ***
RegionScotland  0.04536    1.95643   0.023  0.982
RegionSouth East 13.07848    1.67704   7.799 6.48e-14 ***
RegionSouth West 14.59281    1.88812   7.729 1.04e-13 ***
RegionWales    14.25621    2.17046   6.568 1.73e-10 ***
RegionWest Midlands 21.22315    1.98926  10.669 < 2e-16 ***
RegionYorkshire and The Humber 19.55848    2.20125   8.885 < 2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Residual standard error: 7.886 on 369 degrees of freedom
Multiple R-squared:  0.4419,    Adjusted R-squared:  0.4268 
F-statistic: 29.22 on 10 and 369 DF,  p-value: < 2.2e-16
```

With London as the baseline, `RegionScotland` is the difference between Scotland and London, 0.05 points ($t = 0.02$). The two regions had almost the same mean Leave share (39.09 and 39.14).

Dummy Variables

Two Categories

Multiple Categories

Two Dummy Variables and Their Interaction

Regression with Two Dummy Variables

- Consider voters in the 2024 US presidential election, with $Y_i = 1$ if voter i voted for Trump and $Y_i = 0$ if for Harris, and two dummy variables:

$$X_{White,i} = \begin{cases} 1 & \text{if voter } i \text{ is white (non-Hispanic)} \\ 0 & \text{if voter } i \text{ is non-white (including Hispanic)} \end{cases}$$

$$X_{Col,i} = \begin{cases} 1 & \text{if voter } i \text{ has a 4-year college degree or more} \\ 0 & \text{if voter } i \text{ does not} \end{cases}$$

- Let's regress the Trump vote on these using: $Y = \beta_0 + \beta_{White} X_{White} + \beta_{Col} X_{Col} + u$

Source: Cooperative Election Study 2024 (Schaffner, Shih, Ansolabehere & Pope; doi:10.7910/DVN/X11EP6): 42,028 respondents who reported voting for Trump or Harris; every estimate is weighted by the survey weight.

Example: Trump Vote on White and College

```

R Code
> summary(lm(trump ~ white + college, data = ces, weights = commonpostweight))
~~~~~
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.389210   0.004480   86.87  <2e-16 ***
white        0.239831   0.005130   46.75  <2e-16 ***
college     -0.115424   0.004912  -23.50  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

What does the model imply for the mean of Y (the Trump share) in each category?

	$X_{Col} = 1$	$X_{Col} = 0$	$E[Y X_{Col} = 1, X_{White}]$ $- E[Y X_{Col} = 0, X_{White}]$
$X_{White} = 1$	$\beta_0 + \beta_{White} + \beta_{Col}$	$\beta_0 + \beta_{White}$	β_{Col}
$X_{White} = 0$	$\beta_0 + \beta_{Col}$	β_0	β_{Col}
$E[Y X_{White} = 1, X_{Col}]$ $- E[Y X_{White} = 0, X_{Col}]$	β_{White}	β_{White}	

Example: Trump Vote on White and College

```

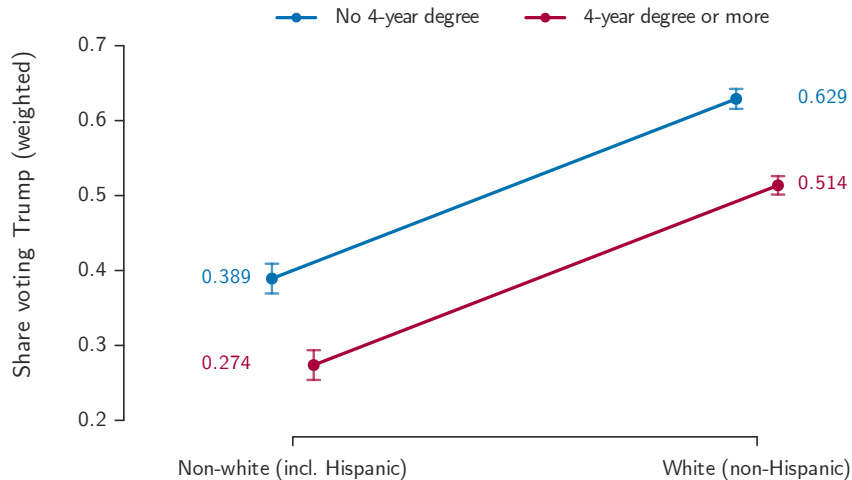
R Code
> summary(lm(trump ~ white + college, data = ces, weights = commonpostweight))
~~~~~
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.389210   0.004480   86.87  <2e-16 ***
white        0.239831   0.005130   46.75  <2e-16 ***
college      -0.115424   0.004912  -23.50  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

What does the model imply for the difference in the Trump share between voters **with** and **without a college degree**?

	$X_{Col} = 1$	$X_{Col} = 0$	$E[Y X_{Col} = 1, X_{White}]$ $- E[Y X_{Col} = 0, X_{White}]$
$X_{White} = 1$	$\beta_0 + \beta_{White} + \beta_{Col}$	$\beta_0 + \beta_{White}$	β_{Col}
$X_{White} = 0$	$\beta_0 + \beta_{Col}$	β_0	β_{Col}
$E[Y X_{White} = 1, X_{Col}]$ $- E[Y X_{White} = 0, X_{Col}]$	β_{White}	β_{White}	

Example: Trump Vote by Race and College Degree



Interaction with Two Dummy Variables

- Now consider our two dummy variables:

$$X_{White,i} = \begin{cases} 1 & \text{if voter } i \text{ is white (non-Hispanic)} \\ 0 & \text{if voter } i \text{ is non-white (including Hispanic)} \end{cases}$$

$$X_{Col,i} = \begin{cases} 1 & \text{if voter } i \text{ has a 4-year college degree or more} \\ 0 & \text{if voter } i \text{ does not} \end{cases}$$

and an interaction term:

$$X_{WhiteXCol,i} = X_{White} \cdot X_{Col} = \begin{cases} 1 & \text{if voter } i \text{ is white and has a college degree} \\ 0 & \text{otherwise} \end{cases}$$

- Let's regress the Trump vote on these using: $Y = \beta_0 + \beta_1 X_{White} + \beta_2 X_{Col} + \beta_3 X_{WhiteXCol} + u$

Example: Trump Vote on White and College

```
_____ R Code _____  
> summary(lm(trump ~ white + college + white_x_college, data = ces,  
             weights = commonpostweight))  
~~~~~  
Coefficients:  
             Estimate Std. Error t value Pr(>|t|)  
(Intercept)   0.370313   0.005058  73.213 < 2e-16 ***  
white          0.268964   0.006280  42.827 < 2e-16 ***  
college       -0.053066   0.009188  -5.775 7.73e-09 ***  
white_x_college -0.087263   0.010869  -8.028 1.01e-15 ***  
---  
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

How to interpret these coefficients?

With survey weights, `lm()` treats the weights as if they measured precision, so its standard errors are too small here. Robust standard errors (Lecture 6-3) are about 1.9 times as large, and the interaction's t is -4.3 .

Worksheet Q2

Let X_{White} indicate a white (non-Hispanic) voter and X_{Col} a voter with a 4-year college degree. Estimate

$$Y = \beta_0 + \beta_1 X_{White} + \beta_2 X_{Col} + \beta_3 (X_{White} \cdot X_{Col}) + u.$$

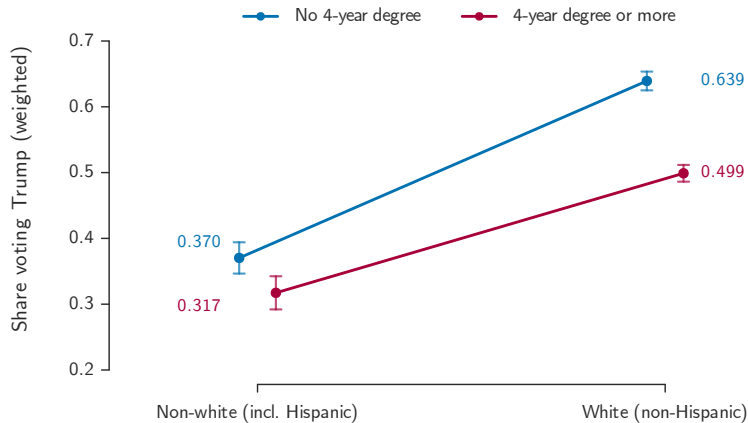
❶ Compute $E[Y|X_{White} = w, X_{Col} = c]$ for $(w, c) \in \{0, 1\} \times \{0, 1\}$.

❷ Show that β_3 equals the difference-in-differences:

$$\beta_3 = (E[Y|1, 1] - E[Y|1, 0]) - (E[Y|0, 1] - E[Y|0, 0]).$$

❸ Give a test for “no interaction” and interpret it.

Example: Trump Vote by Race and College Degree



Among white voters, the college gap is $0.499 - 0.639 = -0.140$. Among non-white voters, it is $0.317 - 0.370 = -0.053$. The difference in differences is $-0.140 - (-0.053) = -0.087 = \hat{\beta}_3$.

Two Dummies and Their Interaction: Dummies for Four Groups

- X_{White} and X_{Col} split the voters into four groups. The interaction term is itself a dummy: it is 1 only for white voters with a college degree.
- So the interaction model is equivalent to a regression on the dummies of one categorical variable with $m = 4$ categories, with non-white voters without a degree as the baseline:

$$Y = \gamma_0 + \gamma_1 D_{White, \text{ no Col}} + \gamma_2 D_{Non-white, Col} + \gamma_3 D_{White, Col} + u,$$

with $D_{White, Col} = X_{White} \cdot X_{Col}$, $D_{White, \text{ no Col}} = X_{White}(1 - X_{Col})$ and $D_{Non-white, Col} = (1 - X_{White})X_{Col}$.

- Both models give each group its own mean, so their fitted values are the four sample group means:

Group	Trump share	Interaction model	Four-group dummies
Non-white, no college degree	0.370	$\hat{\beta}_0$	$\hat{\gamma}_0$
White, no college degree	0.639	$\hat{\beta}_0 + \hat{\beta}_1$	$\hat{\gamma}_0 + \hat{\gamma}_1$
Non-white, college degree	0.317	$\hat{\beta}_0 + \hat{\beta}_2$	$\hat{\gamma}_0 + \hat{\gamma}_2$
White, college degree	0.499	$\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}_3$	$\hat{\gamma}_0 + \hat{\gamma}_3$

- In general, for two categorical variables with k and m categories and every combination observed: an intercept, their $k - 1$ and $m - 1$ dummies and all $(k - 1)(m - 1)$ products give the same fit as an intercept and $k \times m - 1$ dummies for the combined categories.

Dealing with Interaction Terms in R

- In R, multiplicative terms are denoted by “X:Z” in a formula:

```
R Code  
> summary(lm(trump ~ white + college + white:college, data = ces,  
             weights = commonpostweight))
```

~~~~~  
Coefficients:

|               | Estimate  | Std. Error | t value | Pr(> t ) |     |
|---------------|-----------|------------|---------|----------|-----|
| (Intercept)   | 0.370313  | 0.005058   | 73.213  | < 2e-16  | *** |
| white         | 0.268964  | 0.006280   | 42.827  | < 2e-16  | *** |
| college       | -0.053066 | 0.009188   | -5.775  | 7.73e-09 | *** |
| white:college | -0.087263 | 0.010869   | -8.028  | 1.01e-15 | *** |

- You can also use the shorthand “X\*Z” to denote all three terms in one:

```
R Code  
> summary(lm(trump ~ white*college, data = ces, weights = commonpostweight))
```

~~~~~  
Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	0.370313	0.005058	73.213	< 2e-16	***
white	0.268964	0.006280	42.827	< 2e-16	***
college	-0.053066	0.009188	-5.775	7.73e-09	***
white:college	-0.087263	0.010869	-8.028	1.01e-15	***